THE UNIVERSITY *of* EDINBURGH

Title	Computational analyses of A-I RNA editing
Author	Clutterbuck, Daniel Richard.
Qualification	PhD
Year	2006

Thesis scanned from best copy available: may contain faint or blurred text, and/or cropped or missing pages.

Digitisation notes:

- Page numbers 120 & 121 are omitted in pagination, content is continuous
- Pagination in appendix 1 is irregular

Computational Analyses of A-I RNA Editing

Daniel Richard Clutterbuck

Thesis Submitted for the Degree of PhD at the University of Edinburgh

November, 2005

Supervisor: Dr. Colin Semple

If we knew what it was we were doing, it would not be called research, would it? **– Albert Einstein**

Declaration

This thesis has been composed by Daniel Richard Clutterbuck. The author has performed all work within this thesis unless otherwise stated. This work has not been submitted for any other degree or professional qualification.

Acknowledgements

After three years of studying for this PhD there are many people who have helped me along the way. Most importantly, Colin has provided constant support and advice almost every day, ranging from Perl one-liner tips to career advice while wandering round the botanical gardens. I feel very privileged to have been his student. Gogo has also been very supportive. Many tedious afternoons have been avoided while we exchanged gossip and jokes and I will miss our little chats. I also owe a large part of my sanity to Frances and James, the other members of our little lab. We have had some really good times, both in work and out of work.

Much of my work has been carried out in collaboration with Mary O'Connell's lab. Mary, Liam, Jim, Anne, Angela and Leanne have all really helped me with discussions and laboratory work. I have thoroughly enjoyed working with this group. There are many other people at work that I have become friends with and that have helped me through the last four years, including Tom, Anna, all the other students, Nicky Gray, Ian Jackson (my co-supervisor) and the rest of the West Wing. Between them, they have made the HGU a great place to work. I'd like to think that my Andy White impression will be fondly remembered for years to come.

This PhD would never have happened, however, if it weren't for the continued support from my family over the last 26 years. They have always been there to pick me up when I'm down (or my bank balance is down) and I am eternally grateful for it. I'm especially thankful to Jonathon, who has showed me that anything is possible.

I must also thank all my friends who have always kept my spirits up (and kept me filled with spirits a fair amount as well). Foremost are Barney, Scouse, and the rest of the Edinburgh University Hill walking Club. I will sorely miss our weekly outings to the pub and our trips to the highlands. Kristy deserves a special mention for keeping me happy for the last year and for helping me explore so much of Edinburgh.

Finally, I would like to make a special mention to Claire, who stood by me for the first three years of my postgraduate career and was always there to look after me.

Contents

Index of Figures

Index of Tables

Abbreviations

The following abbreviations are used repeatedly in this text.

A-I	Adenosine to inosine
ADAR	Adenosine Deaminase Acting on RNA
ADAT	Adenosine Deaminase Acting on tTRNA
AMPA	Alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
APOBEC	Apolipoprotein B mRNA-editing Enzyme Catalytic polypeptide
ALS	Amyotrophic Lateral Sclerosis
cDNA	Complementary DNA
C-U	Cytosine to uridine
DSH	Dyschromatosis Symmetrica Hereditaria
dsRBD	Double-stranded RNA binding domain.
dsRNA	Double-stranded ribonucleic acid.
ECD	Edited Complementary Duplex Prediction
ECS	Edited Complementary Sequence
ES	Edited Sequence
EST	Expressed Sequence Tag
GluR-A to D	Ionotropic Glutamate/AMPA receptor subunits A, B, C or D.
GluR-5/6	Kainate Receptor Subunits.
$5HT_{2C}R$	Serotonin Receptor (type 2C)
LOD	Log of Odds
mRNA	Messenger RNA
PCR	Polymerase Chain Reaction
RT-PCR	Reverse Transcriptase PCR
SNP	Single Nucleotide Polymorphism
UTR	Untranslated region.

Publications Arising

POCUS: Mining Genomic Sequence Annotation to Predict Disease Genes.

Frances S Turner, Daniel R Clutterbuck and Colin AM Semple.

Genome Biology 2003, 4:R75.1-9

See *Appendix 2* for details.

A Bioinformatic Screen for Novel A-I RNA Editing Sites Reveals Recoding Editing in BC10.

D.R. Clutterbuck, A. Leroy, M.A. O'Connell and Colin AM Semple.

Bioinformatics 2005, 21.11:2590-2595

See *Chapter 3* for details.

The Transcriptional Landscape of the Mammalian Genome

The Fantom Consortium and Riken Genome Exploration Research Group and Genome Science Group (Including DR Clutterbuck & Colin Semple)

Science 2005, 309:1559-1563

See *Appendix 3* for details.

Using Comparative Genomics & RNA Structure to Find RNA Editing Sites

Daniel Clutterbuck (Poster Presentation from BioSysBio conference, 2005)

BMC Bioinformatics 2005, 6:Suppl 3:P6

See *Chapter 4* for details.

1 Introduction

1.1 RNA Editing

RNA editing is any modification of RNA that is distinct from RNA splicing, capping or 3' processing[1]. These events can consist of ribonucleotide insertion, deletion or conversion. The results of these modifications have been shown to affect protein diversity by introducing altered splicing patterns, frame shifts, alternative start or stop codons, or directly affecting the protein sequence. More recently it has become apparent that many editing sites are situated in UTRs2, suggesting that they mediate a regulatory function. Editing is a widespread phenomenon that has been identified in numerous species including human, mouse, rat1, fish[3], nematode[4], fruit fly[5], plants[6], protozoa[7] and bacteria[8]. In bacteria, however, this process is termed modification. The two best-studied forms of RNA editing in higher eukaryotes are adenosine to inosine events (A>I), or cytosine to uridine (C>U) events. A-I editing appears to be the most common[9]. The projects described in this thesis are only concerned with A-I RNA editing. The broader topic of RNA editing has recently been reviewed[1,10].

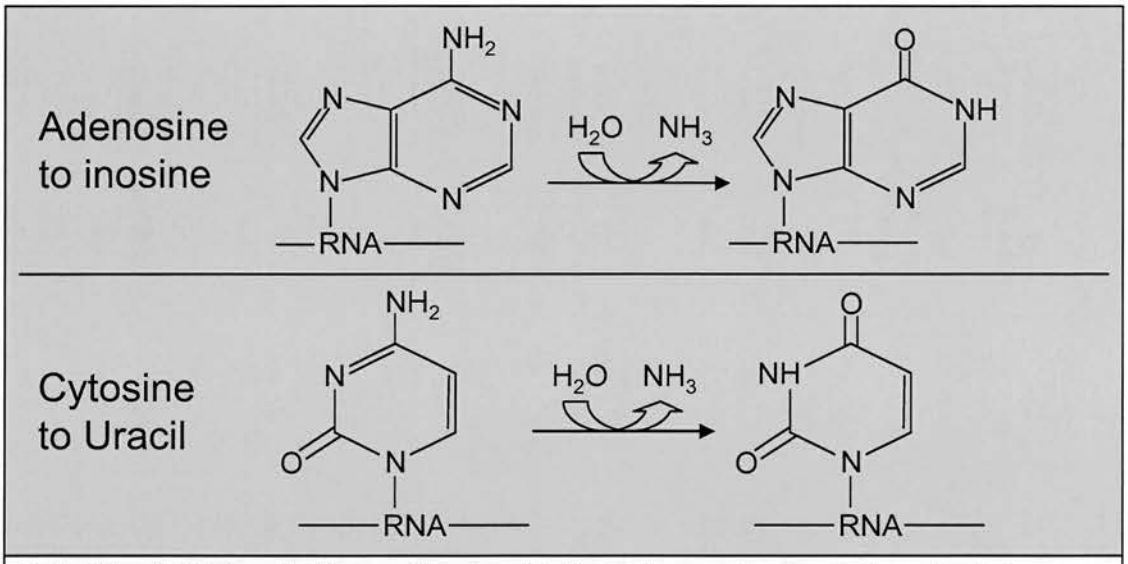
1.1.1 What Is A-I RNA Editing?

The process of A-I RNA editing involves the hydrolytic deamination of adenosine at the C6 site, resulting in a conversion to inosine. Figure 1.1 (p2) shows the deamination reactions that occur in A-I and C-U editing. If an inosine is located within coding sequence then the translational machinery will read this as a guanosine, which could result in a re-coding of the protein sequence[1]. For example, if the codon 'AGA' were edited to 'AGI', then this would be translated as if it was a 'AGG' codon. This would result in a change in the encoded amino acid from an arginine to a glycine.

1.1.2 ADATs & ADARs

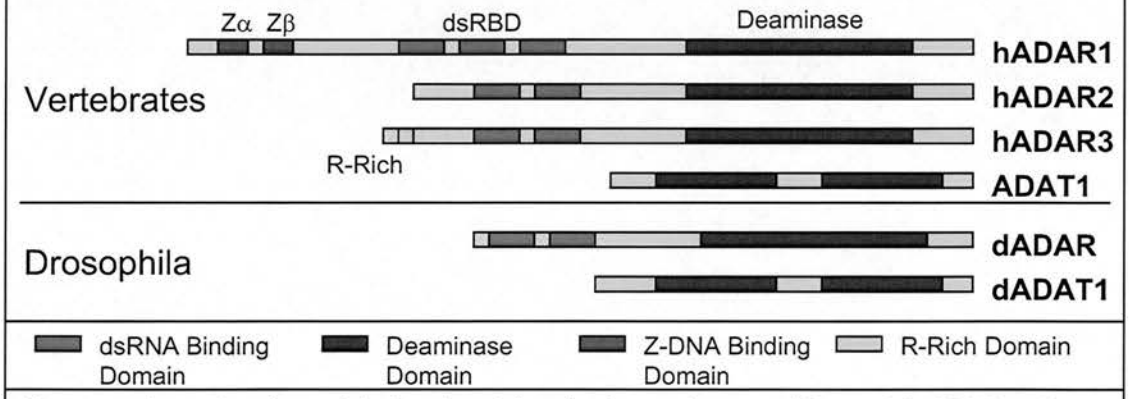
The conversion of adenosine to inosine is mediated by a family of enzymes called ADARs (Adenosine Deaminases that Act on RNA). These genes appear to have evolved from ADATs (Adenosine Deaminases that Act on tRNAs)[11-13], which are

Figure 1.1. The Chemistry of A-I & C-U Deamination



Both A-I and C-U deamination reactions result in the exchange of an ammonia molecule for a water molecule. This is termed a hydrolytic deamination. The deaminase domains of the APOBECs and the ADARs appear to be of common origin, which could account for the similarity of the two reactions.

Figure 1.2. The Domain Structure of ADARs & ADATs



The approximate locations of the functional domains in vertebrate and *Drosophila* ADAR and ADAT genes. Each ADAR gene contains two or more dsRNA binding domains, which enable them to bind RNA duplexes. The ADATs do not contain dsRBDs. The deaminase domain is shared by all of these genes, but is split into two in the ADATs. ADAR1 contains two Z-DNA binding domains and ADAR3 contains an R-rich domain. (Maas et al, 2003, JBC).

enzymes that deaminate adenosines in tRNAs. Historically the conversion of adenosine to inosine by ADATs within tRNAs is termed modification, while the same change mediated by the ADARs in mRNA is termed editing. Figure 1.2 (p2) shows the domain structures of these genes. In humans the ADAR family has three known members, *ADAR1*, *ADAR2* and *ADAR3*, the first two of which are ubiquitously expressed[14] and the third is brain specific[15]. As yet there are no known targets for *ADAR3*[16], whereas the other two enzymes have a number of known targets, some of which can be edited by both enzymes *in vitro*[17,18]. ADARs contain two or three dsRNA binding domains and a deaminase domain. The ADARs are thought to function as homodimers[19,20]. It is believed that once a potential editing site has been bound, the target base is physically flipped out for the deamination reaction[21]. This process is thought to be similar to that of DNA methyl transferases. Importantly, this process does not require co-factors *in vitro*, which suggests that much of the specificity is provided by a direct interaction with the RNA.

There are many different splice variants for these enzymes, which allow for diversity in target specificity and efficiency[11]. Indeed, the rat *ADAR2* enzyme auto-regulates itself by editing its own pre-mRNA transcripts[22]. This results in the creation of a new splice site, which gives rise to an alternative splice form with a shifted open reading frame and results in a non-functional protein. The *Drosophila dADAR* gene also edits its own transcript[5]. This also down-regulates the activity of the enzyme, but by changing an amino acid in the deaminase domain.

The efficiency of A-I RNA editing is rarely 100% in mammals, with the only exception being an editing site that results in an Q to R amino acid conversion in the Glutamate Receptor B (*GluR-B*) transcript[23]. This edited site is called the *GluR-B* Q/R site, and demonstrates the general nomenclature for naming known edited sites (gene then the coding change). The editing efficiency varies widely both between the different sites and between tissue and developmental stages. For example, the serotonin ($5HT_{2C}R$) receptor transcript has five edited sites in close proximity. Each of these sites is edited at different frequencies, and these frequencies vary in tissue-specific patterns[24]. The result is that there are 12 principal isoforms in the brain, each with a different expression pattern. Interestingly, the proportions of these isoforms are altered in suicide victims, while Prozac™ appears to have a balancing effect on these proportions in mice[25]. Each isoform also incurs different sensitivity to lysergic acid (LSD)[26]. The *cacophony* gene in *Drosophila* is potentially even more variable, with ten edited sites, which could result in over 1,000 isoforms by editing alone[27,28].

1.1.3 AMPA Receptors

A brief description of AMPA receptors is required to fully discuss the following observations. AMPA receptors mediate the majority of fast excitatory synaptic transmission[29]. They are composed of four subunits, *GluR-A,B,C & D*, which combine to form tetramers with a central ion channel[30]. The receptors can be comprised of just one type of subunit, or can be heteromeric. If the AMPA receptor does not contain a *GluR-B* subunit that is edited at the Q/R site, then it will be permeable to sodium, potassium and calcium. However, if the AMPA receptor contains a GluR-B subunit that is edited at the Q/R site then the calcium permeability of the resultant AMPA receptor becomes negligible[31]. In contrast editing at the *GluR-B,C&D* R/G sites hasten the rate at which AMPA-type channels recover from desensitisation[32]. Each of the four subunits can be alternatively spliced to contain either a 'flip' exon, or a 'flop' exon. These are short adjacent exons of 115bp that differ by only a few amino acids. However, for each subunit the splice variants have quite different desensitisation kinetics[33]. The combination of variability in subunit composition, splice variants and editing variants results in a wide variety of resultant AMPA channels.

1.1.4 ADAR Mutations in Model Organisms

Although it is widely understood that the ADARs diverged from the ADAT family of enzymes, it is not clear what function they originally served. One way to answer this question is to observe what happens when transgenic animals are generated that are deficient in ADAR activity. Table 1.1 (p5) shows a summary of these phenotypes. In the fruit fly, *Drosophila melanogaster*, there is only one ADAR gene, *dADAR*. The null mutant of this gene has a normal lifespan, but has brain lesions and serious behavioural deficits[34]. A more severe phenotype is shown by the homozygous mouse null mutant for the *ADAR2* gene. The mice died soon after weaning and were prone to epileptic seizures. Interestingly, the impaired phenotype appeared to result entirely from under-editing of the *GluR-B* Q/R edited site. The phenotype reverted to wild type when both alleles for this transcript were substituted with alleles encoding the edited version[35]. This suggests that the primary purpose of *ADAR2* in the mouse is to edit this transcript.

Table 1.1. The Phenotypes of ADAR Deficient Animals

Organism	Genotype	Affected Tissue	Transcript	Phenotype
Nematode	*ADAR*-1/2$^{-/-}$	Unknown	Unknown	Problems with chemotaxis[4].
Fruit fly	*dADAR*$^{-/-}$	CNS	Unknown	Normal lifespan. Behavioural and locomotive problems[5].
Mouse	*ADAR1*$^{-/-}$	Many Tissues	Unknown	Widespread apoptosis[36], esp. in the liver[36], and death by E11.5.
Mouse	*ADAR1*$^{+/-}$	Blood	Unknown	Died before E14 due to defects in the haematopoietic system[37].
Mouse	*ADAR2*$^{-/-}$	CNS	*GluR-B* (Q/R)	Epileptic seizures and death by P20[35].
Human	*ADAR1*$^{+/-}$	Skin	Unknown	Dyschromatosis Symmetrica Hereditaria (DSH). A skin pigmentation disease[38].
Human	Unknown	Motor Neurons	*GluR-B* (Q/R)	Reduced editing of Q/R site correlates with motor neuron death & amyotrophic lateral sclerosis (ALS)[39].

Initially, this phenotype was thought to be entirely due to increased calcium permeability of the glutamate (also termed AMPA) receptors, mediated by editing of the Q/R site within the *GluR-B* pore loop region[40]. However, Greger *et al* showed that editing of this site is also required for the correct assembly of the glutamate receptor subunits[41]. Editing at the Q/R site retains the *GluR-B* subunit in the endoplasmic reticulum, until it has bound three other glutamate receptor subunits to form a complete receptor. The unedited version is not retained and readily tetramerises, resulting in increased Ca^{2+} permeability in synapses[41].

Two groups have created homozygous null mutants of the *ADAR1* gene. In both cases the mice died before E12.5 and showed a phenotype consistent with widespread apoptosis[36], especially in the liver[42]. Heterozygous null *ADAR1* mutant mice were also generated, which died before E14 due to defects in the haematopoietic system[37].

In humans there have been two diseases that have been identified that are strongly involved in the function of the ADARs. It has been suggested that Amyotrophic lateral sclerosis is caused by inefficient editing of the *GluR-B* Q/R site, as a strong correlation is seen between this and the death of spinal motor neurons[39]. It seems plausible that mis-regulation of *ADAR2* might underlie this disease, however, further experiments are required. More direct evidence identified mutations in *ADAR1* as the causative mutations behind a skin pigmentation disease called Dyschromatosis Symmetrica Hereditaria (DSH)[38]. Two of the initially reported mutations resulted in a truncated protein, one was thought to alter the catalytic site and the final mutation resulted in a coding change of unknown consequence. This phenotype is remarkably different from the null mutant in the mouse. It is not clear why this difference exists. Many more mutations in this gene have been found in patients with DSH[43-47].

1.1.5 C-U RNA Editing by APOBEC Enzymes

The conversion of cytidine to uridine residues is carried out by cytidine deaminases, a family of enzymes that includes *APOBEC1, APOBEC2, APOBEC3* and *AID*[1]. *AID* has been shown to mediate immunoglobulin class switch recombination[48]. *APOBEC1*, the best studied cytidine deaminase, edits apolipoprotein B (*APOB*) mRNA to create a premature stop codon[49]. This gives rise to a shortened version of *APOB*, which affects the regulation of *LDL* (Low density lipoprotein). This has medical importance, as high levels of LDL cholesterol constitute one of the main risk factors in coronary heart disease. *APOBEC1* requires an auxiliary factor, *ACF* (*APOBEC1* Complementation Factor), to efficiently edit *APOB*[50]. There are many similarities between the C-U editing reaction and the A-I editing reaction. The APOBECs and the ADARs contain a homologous deaminase domain. The targeting of the ADARs to dsRNA is thought to result from their dsRNA binding domains, while the targeting of the APOBECs is thought to result from the dsRNA binding domains in the *ACF* cofactor. In addition to editing, this complex has been shown to suppress nonsense-mediated decay[49]. Near the completion of this thesis, another C-U editing site was identified in a glycine receptor[51].

1.1.6 Other Types of Editing

In addition to A-I and C-U editing in vertebrates there are many other well-studied forms of RNA editing. C-U editing has been widely described in plant mitochondria and chloroplasts[52]. Viruses also use editing, such as the Hepatitis Delta Virus, which requires A-I editing by its host in order to complete its life cycle[53]. There are also several species that show widespread insertion or deletion editing sites. These include uridine insertion/deletion in kinetoplastid protozoa mitochondria (trypanosomes)[54], cytosine or dinucleotide insertion in *Physarum* mitochondria (slime moulds)[55,56], and guanosine insertion in viruses[57].

There is also limited evidence for additional types of editing in vertebrates. Recently, a putative U-C event was described in a mouse mitochondrial transcript[58]. This type of editing is considered unlikely by some as the evidence for this site is poor and the proposed addition of an amino group is energetically very expensive. In addition to this site, there are several other unconfirmed sites that are neither A-I or C-U[59-62].

1.2 A-I Edited Sites

1.2.1 A Wide Range of A-I Edited Sites

ADARs require their substrates to be double-stranded RNA[63]. The specificity of the ADAR enzymes results from interactions between this dsRNA and the dsRNA binding domains in the enzyme.If a substrate is a perfectly base-paired duplex longer than 50bp, then the substrate is edited until roughly half of the adenosines are deaminated[10,64]. There is still some selectivity, however, as it tends to be the same adenosines that are edited. This form of editing, as seen with the 4f-rnp transcript, is termed hyper-editing[65]. Recently, it has been shown that hyper-edited dsRNAs are cleaved in the cytoplasm[66,67]. This process requires *Tudor-SN*, a vital component of the RNA-induced silencing complex (RISC). These observations support the possibility that hyper-editing, in addition to RNAi, may be a form of defence against viruses that replicate through a double-stranded RNA intermediate.

In humans there is another class of edited sites, which result from inverted copies of Alu repeat sequences[68-70]. The Alu sequences base pair, often resulting in very long duplexes. Editing of these duplexes occurs at a high proportion of the adenosines, although the editing frequency is often very low (typically less than 5%). It is not clear if any of these sites are functional; although some have been shown to result in alternative splicing and exonisation of Alu repeat sequences. Editing of this type can also be observed in the mouse and between other types of repeat, however, the number of edited sites is greatly diminished[71]. This is possibly due to the fact that Alu repeats are exceptionally common in the human transcriptome, but related repeats in mouse are rare. This means that there is a greater probability of finding two Alu repeats near to each other in the same transcript, which would provide an A-I editing substrate.

In some cases there is evidence that the dsRNA is formed by an intermolecular reaction. Recent advances have shown that sense-antisense pairs of transcripts are common in mammalian genomes[72,73]. I was part of the FANTOM consortium that published one of these papers (see Appendix 3). It is possible that duplexes formed between the sense and antisense transcripts could be good targets for A-I editing, however, with the exception of the 4f-rnp transcript in *Drosophila*, there is no

evidence to support this[65]. Complementary RNAs can also regulate editing, such as with the *5HT$_{2C}$R* receptor[74,75]. In this case, a snoRNA (small nucleolar RNA) has been shown to mediate the modification of the putative editing site by 2'O-methylation, which can reduce the rate of deamination by 200-fold *in vitro*[76].

Most of the well-studied edited sites, however, result in editing of specific adenosines within highly conserved and relatively short imperfect RNA duplexes formed through intra-molecular base pairing. These duplexes vary in length from just under 30bp to over 120bp (Figure 4.1 – p75). It is thought that the bulges and loops are required to provide the specificity to guide the enzyme to the specific adenosines[77]. This type of editing is referred to as 'specific editing'. The majority of these sites have attributable functions, such as recoding proteins or affecting splicing. However, some occur in non-coding regions of transcripts and their functions are unknown. For the purposes of this thesis, the edited sites that result in a specific protein recoding event were considered the most interesting group of sites to study as they are the most likely to be functional. For this reason, this thesis is mainly concerned with this type of site.

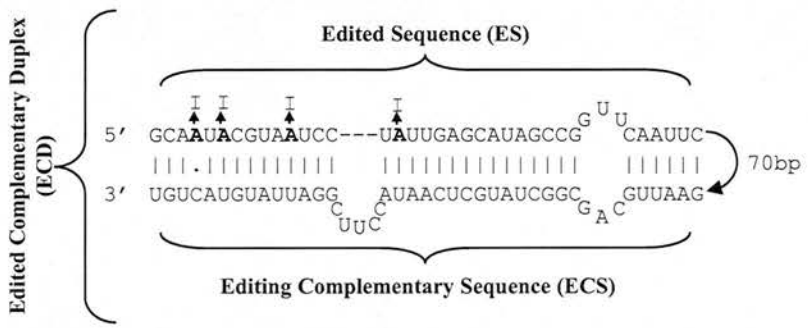
1.2.2 Anatomy of an A-I editing site

Figure 1.3 (p10) shows the essential elements of the serotonin (*5HT$_{2C}$R*) receptor edited sites. Section A shows a typical edited dsRNA duplex structure, with 70bp of intervening sequence. The duplex is both relatively short and relatively imperfect (i.e. it contains some non-complementary bases). Section B shows the exonic location of the edits and demonstrates the high level of sequence conservation associated with this cluster of edited sites. Each of the four edited sites shown results in re-coding of an amino acid in the resultant protein.

When describing these structures it is imperative that strict terminology is used. Some of the terminology given here is novel, as there was no existing terminology that could sufficiently explain the given features. The most important terms are described below;

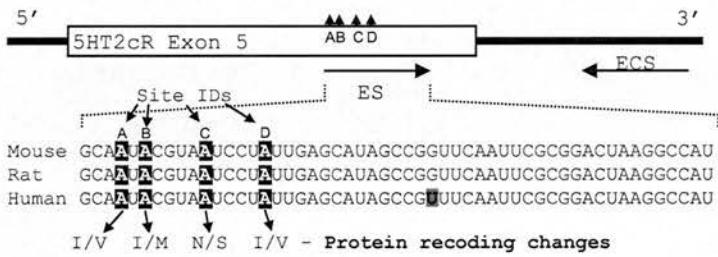
Figure 1.3. A Typical Recoding Edited Site

A. Inverted Repeat Structure



The predicted RNA structure of the $5HT_{2C}$ receptor edited sites. The edited nucleotides are marked with an 'I', which indicates the change to inosine. The 5 prime and 3 prime ends of this sequence are indicated. The structure is annotated with formal terms for describing these structures. The edited sequence (ES) and the editing complementary sequence (ECS) are separated by 70bp.

B. Overview and Sequence Conservation

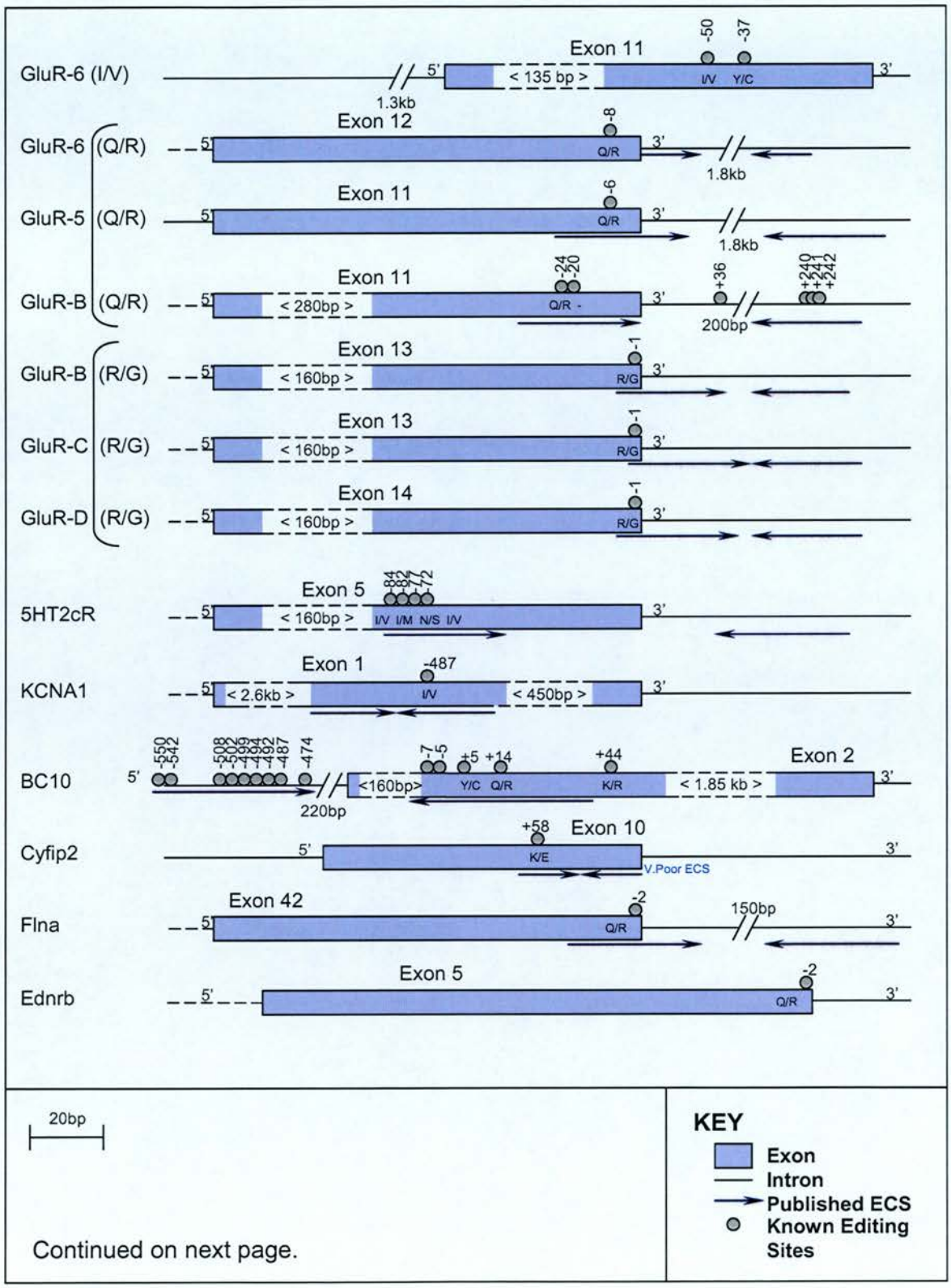


The location of the inverted repeat structure described in section A in relation to the exon structure of the gene. A section of the alignment is also shown. This demonstrates the high levels of sequence conservation that can be observed across the edited sequence (ES). The only nucleotide that is not completely conserved in mouse, rat & human is highlighted in red. The recoded amino acids have been indicated for each site (unedited/edited).

- **RNA Duplex** – A double-stranded RNA structure formed through base pairing of two sequences. Typically, this will result from two complementary regions of the same transcript, although these will not necessarily be adjacent regions. RNA duplexes may also occur between separate RNA molecules. An RNA hairpin is a special case of RNA duplex where the two base-pairing regions have little or no intervening sequence.
- **Inverted Repeat** – This is when a copy of a sequence is found in the same molecule, but in the opposite direction on the same strand. This can occur in both DNA and RNA. If the situation arises in RNA this can give rise to a perfect or near perfect duplex, depending on the similarity of the two sequences. Transcripts with inverted Alu repeats are particularly common in humans[71].
- **Edited Sequence (ES)** – This is the half of the duplex that contains the edited nucleotides or if no edited nucleotides are known, it is the half that typically occurs in an exon.
- **Editing Site Complementary Sequence (ECS)** – This is the half of the duplex that does not contain the edited nucleotides. The ECS is typically located in the adjacent intron. This is an established term in the field[1].
- **Editing Complementary Duplex (ECD)** – This describes the entire duplex structure, incorporating both the Edited Site (ES) and the Editing Complementary Site (ECS), but not the intervening sequence. **Unless stated otherwise, all ECDs referred to in this thesis are only *predicted* ECDs.**

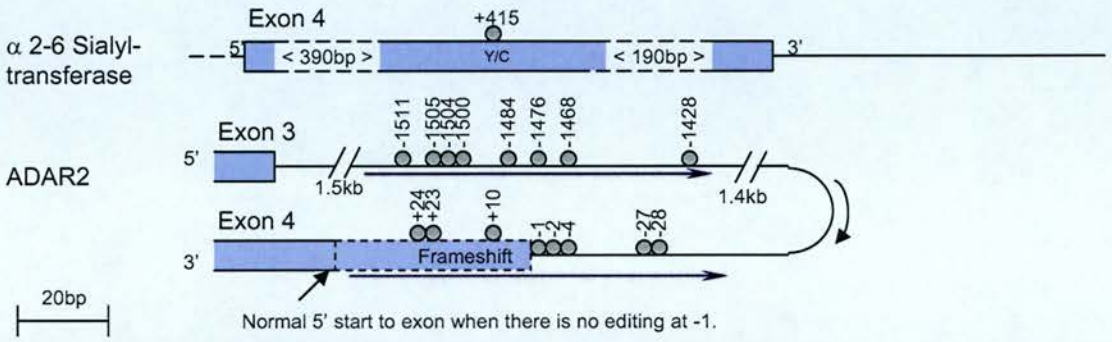
Many of the features shown for the *5HT$_{2C}$R* receptor are common to the majority of the other known edited sites. Figure 1.4 (p12) shows the vertebrate edited sites that result in re-coding proteins, either directly or through altered splicing, including the *5HT$_{2C}$R* site. Most, if not all, of the A-I edited sites are widely assumed to be mediated by ECDs where both the ES and the ECS occur in the same transcript, however, not all these sites have published ECDs. Published ECD predictions have been shown where available. The references and the experimental support for each of these ECDs are shown in Table 1.2 (p14), which lists mammalian genes that have been shown to result in protein recoding. The distance between the ECD halves varies between zero for the *GluR-B* R/G site[32] and 1.8Kb for the *GluR-6* Q/R site[78]. The length of these ECDs vary, with most of them being between 20 and 30bp. However,

Figure 1.4. The Known Mammalian Protein Recoding Edited Sites



Continued on next page.

Figure 1.4. The Known Mammalian Protein Recoding Edited Sites (Continued)



The 15 known protein recoding edited sites, showing the locations of the individual sites and their associated known ECSs. All transcripts are shown in the 5' to 3' direction and are drawn approximately to scale. The three glutamate receptor Q/R regions are grouped together due to their similarity, as are the three R/G regions. For the longer exons, gaps are shown with the intervening length given in the gap. Similarly, long intronic regions are denoted by two parallel slashes and an associated length. Where the protein is modified by editing, the amino acid change is indicated. The number above each site specifies the exact location of each site. Coordinates are relative to the nearest exon start/end. ADAR2 is alternatively spliced; editing at the –1 site creates an alternative splice site, resulting in the inclusion of 47bp and a frame shift. The resultant protein is not thought to be functional. The additional 47bp are indicated by the dashed line surrounding the initial segment of this exon. As well as giving an overview of the known sites, this figure demonstrates that mammalian recoding editing sites often occur in clusters and tend to be near the ends of exons.

KEY

- **Exon**
- **Intron**
- **Published ECS**
- **Known Editing Sites**

Table 1.2. Previous Evidence for ECS Structures in the Known Edited Sites

ECS Type	Known Edited Site	Psuedonym	Ensembl Gene ID	Ensembl Exon ID	ECS Reference (or original reference)	ECD Evidence
Experimentally Validated	GluR-B R/G	AMPA-2	ENSMUSG00000033981	ENSMUSE00000223221	Lomeli, 1994[32]	Nucleotide substitutions & rescue in GluR-B R/G site only.
	GluR-C R/G	AMPA-3	ENSMUSG00000001986	ENSMUSE00000478281	Lomeli, 1994[32]	
	GluR-D R/G	AMPA-4	ENSMUSG00000025892	ENSMUSE00000519849	Lomeli, 1994[32]	
	GluR-B Q/R	AMPA-2	ENSMUSG00000033981	ENSMUSE00000223233	Higuchi, 1993[23]	Minigene nuclelotide substitution & rescue.
	GluR-5 Q/R	Grik1	ENSMUSG00000022935	ENSMUSE00000515520	Herb, 1996[78]	A series of deletions.
	GluR-6 Q/R	Grik2	ENSMUSG00000056073	ENSMUSE00000495901	Herb, 1996[78]	Deletion series. Nucleotide substitution & rescue.
	5HT		ENSMUSG00000041380	ENSMUSE00000250405	Burns, 1997[24]	Nucleotide substitutions & rescue
	BC10	Blcap	ENSMUSG00000057453	ENSMUSE00000472446	Clutterbuck, 2005[79]	Inosines in intronic sequence.
	ADAR2	AdarB1, Red1	ENSMUSG00000020262	ENSMUSE00000102704	Dawson, 2004[63]	Nucleotide substitution & rescue. (also sequence conservation lower in bulges).
	KCNA	Kv1.1	ENSMUSG00000047976	ENSMUSE00000453352	Bhalla, 2004[80]	Nucleotide substitutions & rescue
Prediction Only	Flna	-	ENSMUSG00000031328	ENSMUSE00000486137	Levanon, 2005[81]	MFold Prediction - Good
	Cyfip2	-	ENSMUSG00000020340	ENSMUSE00000103611	Levanon, 2005[81]	MFold Prediction - Poor
No prediction	GluR-6 I/V	Grik2	ENSMUSG00000056073	ENSMUSE00000492643	(Schmauss, 2002[30])	No prediction
	Ednrb	-	ENSMUSG00000022122	ENSMUSE00000123754	(Tanoue, 2002[82])	No prediction
	Alpha2,6-Sialyltransferase	-	ENSMUSG00000022885	ENSMUSE00000376978	(Ma, 1997[83])	No prediction

The known edited sites that result in protein recoding are described here. They have been split into three categories, depending on the degree of evidence for their publshed ECD structures, where available.

the *ADAR2* self-editing site ECD is approximately 100bp long[63]. In each case shown here, the ES is either totally within the exon, or more commonly, it adjoins or overlaps one end of the exon. In contrast, all the mammalian ECSs are within an adjacent intron, with the exception of the *KCNA1* site, where both halves of the ECD occur within the exon[80]. This figure also includes an edited site identified by the work in this thesis[79]. This site was also identified by another group[81].

1.2.3 High Sequence Conservation of ECDs

Figure 1.3 (p10) shows the high level of sequence conservation of the *5HT$_{2C}$R* edited site in three mammals. The only nucleotide that is not completely conserved is highlighted. Most of the known recoding edited sites exhibit similarly high levels of conservation. For example, the region containing *GluR-B* Q/R edited site has over 99% nucleotide similarity for a 120bp region conserved between mouse and human (Table 3.3 – p60). The remaining known protein recoding sites are not so exceptionally well conserved, however. The known edited sites and ECSs are sometimes observed across a large range of species. For example, the *GluR-B* R/G ECD can be clearly observed in mouse, rat, human, chicken and fish[85]. This suggests that this edited site arose over 450 million years ago in a common ancestor of fish and mammals[86]. In each of these cases both halves of the ECD are very well conserved[85].

The reasons for such high levels of sequence conservation in ECDs are not entirely clear. Even the most vital coding sequences are often only completely conserved in the first two positions of each codon. The third position of each codon is normally not strongly constrained in coding sequences, and would be expected to vary due to neutral drift over the timescales examined here. As demonstrated by Figure 1.3 (p10), the edited sites are typically conserved in almost all positions. The most plausible explanation for this is that every position in the sequence is required to stay the same in order to generate the correct duplex structure. This may reflect a high level of accuracy required for this structure to recruit and accurately target the correct ADAR isoforms to the correct adenosines. Even so, the same structure could be achieved through compensatory mutations. For example, an A-U base pair could mutate to a G-U or a G-C base pair without changing the overall structure.

There is a precedent in the literature for exceptionally well-conserved sequences[87,88]. These sequences appear to be non-transcribed and their locations are correlated with developmental genes. This suggests that these sequences have some kind of regulatory role, however, it is still unclear how these sequences have remained so unchanged through evolution. The possibilities that have been considered can broadly be split into three categories including locally increased DNA repair, protection from mutation, and selective constraint due to the sequence performing multiple functions[87]. Similar explanations could be applied to explain why the edited sequences are so well conserved, but selective constraint is the most likely.

1.2.4 The Specificity of A-I Editing

The source of the specificity of A-I editing has been and remains an enigma. *In vitro*, the ADARs are able to edit suitable targets without additional co-factors. This suggests that the specificity is a quality of the enzyme itself. The most obvious candidate regions of the protein would be the dsRNA binding domains that each enzyme has. These domains interact with the sugar-phosphate backbone of the RNA, without making direct contact with the bases[89,90]. This would suggest that the only specificity they provide is in the requirement of the target to form an RNA duplex. However, it has been shown the ADAR2 dsRBDs bind selectively to a duplex mimicking the Q/R site[91]. Additional data shows that the dsRBDs of ADAR2 may also contribute to the specificity beyond recognising the duplex, possibly through selectively aiding the flipping out mechanism of the editing reaction[21]. Internal loops and bulges in the RNA duplex have been shown to add specificity[77]. These observations suggest a model in which the dsRBDs target ADAR2 to the correct duplexes, then help in the editing reaction as well.

However, the main determinant of the specificity of RNA editing appears to come from the deaminase domain. In an elegant experiment, the deaminase domains were swapped between ADAR1 and ADAR2. The specificities of these genes overlap, but are not identical. The specificities of the resultant chimaeras matched those of the deaminase domains, not the rest of the protein[92].

An alignment of the targets of *ADAR2* has shown that there is no clear consensus sequence that targets the ADARs to their sites. There are base preferences for the

bases immediately adjacent to the edited adenosine, but even these are not particularly strong[63].

It is possible that the specificity *in vivo* is provided by additional factors. The observation that many of the edited genes are edited in more than one exon suggests that there is some property of the entire transcript that makes it a good editing target. One possibility is that the specificity is provided at the promoters of these genes, such that active ADAR enzymes are only recruited to a subset of genes that require editing. Indeed, one of the ADAR1 isoforms contains Z-DNA binding domains, that have been proposed to target ADAR1 to the site of transcription[93-96]. This question is still under close scrutiny in the A-I RNA editing field. In summary, the specificity of the ADARs appears to come from a combination of sources including the dsRNA binding domains (and their associated RNA duplexes), the deaminase domains and potentially other sources of specificity.

The story for C-U RNA editing by APOBEC1 is far simpler. The specificity is provided by an 11bp mooring sequence, usually five nucleotides downstream of the edited C, which base-pairs with a 3 prime efficiency element[97]. Mutagenesis has also identified a 5 prime efficiency element and a requirement for A+U rich bulk RNA[1]. Using homology to these sequences, a second C-U target was identified, the *NF1* tumour suppressor[98].

1.2.5 The Evolution of A-I Editing

It has been shown that the ADAR enzymes have probably diverged from the ADAT family, through the incorporation of two or more dsRNA binding domains[11]. However, many questions remain about how and why the known editing sites have evolved. One possibility is that the original function of the ADARs was as an anti-viral or anti-retroviral response. This is supported by the fact that one of the isoforms of *ADAR1* is interferon inducible[99,100]. It has been shown that A-I RNA editing interferes with RNAi[101-104] and hyper-edited RNA has been shown to be cleaved via a central component of the RNAi pathway[67,105]. This also suggests a link between editing and viral defence. There are also several reports of A-I editing of viruses in the literature, including measles virus, hepatitis delta virus (HDV), parainfluenza, respiratory syncytial virus, VSV DI particle, polyoma virus, and two avian retroviruses[10]. Some of these involve fully double-stranded targets, while others are

mediated by antisense transcripts or RNA hairpins. Interestingly, the HDV target actually requires editing for replication to occur[106]. There have also been observations of C-U editing of viruses by APOBEC3G[107].

Once the ADAR system had become established, it is plausible to think that it could have been turned to mediate additional functions, such as re-coding of proteins or affecting splicing. In order to do this, however, suitable substrates are required, which in this case must be dsRNAs. There are several possible sources of dsRNAs, including antisense transcription, existing duplex structures and inverted duplications of exon sequences (which are often mediated by retro-transposons). The latter two categories could both have given rise to the edited sites that we observe today.

Once an edited target has become established, the high levels of sequence conservation observed suggest that it is under very strong selective constraint. Some of the genes containing these edited sites then undergo gene duplication and subsequent divergence. This can be seen for the *GluR-B,C&D* R/G edited sites, which clearly maintain an ECD structure in each gene that is almost identical across mammals.

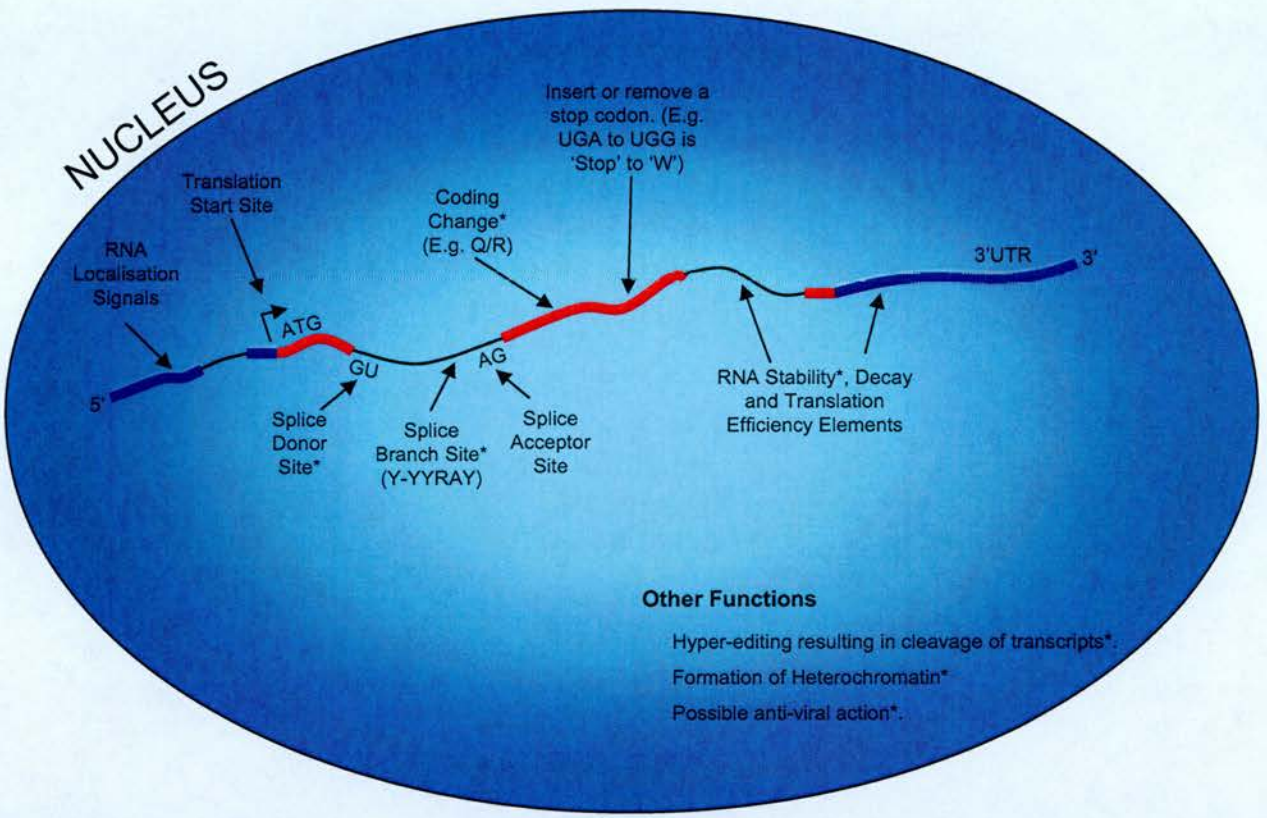
The *Drosophila synaptotagmin-I* (*syt*) gene appears to share some of its edited sites with mosquitoes and butterflies. However, the ECD structures are completely different, with a long-range pseudoknot structure in the fruit fly and a simple imperfect duplex in butterflies[108]. Interestingly, the sea hare, a type of sea slug, produces the same effect through a third mechanism; it has two alternatively spliced exons, one with the normal form and the other with the edited form[108]. It has been suggested that this represents convergent evolution.

1.2.6 The Functions of A-I RNA Editing

The fundamental effect of A-I editing is to introduce inosine nucleotides in place of adenosines. This small change could have a major effect on many systems, many of which have been shown to occur. Some of these possible and known effects have been shown in Figure 1.5 (p19).

As discussed above, there are several reports of hyper-editing of viral RNA. This results in recoding the viral sequences, making them ineffective. While this may

Figure 1.5. Putative and Known Effects of A-I RNA Editing.



Potential functional consequences of A-I RNA editing in a normal protein coding transcript. Many have already been shown to be affected by A-I editing (these are indicated by a *).

KEY

- Untranslated regions (UTRs)
- Protein coding sequence
- Intronic sequence

account for the hyper-editing type of A-I editing, it does not account for more specific editing of transcripts in the host species. Hyper-editing has also been implicated in the formation of heterochromatin although the precise mechanisms are unclear[109].

For the majority of the Alu-repeat mediated editing sites there does not appear to be any clear consequences, except possibly the stabilisation or destabilisation of the duplex formed between the two Alu-repeats[70]. A few of these sites can result in altered splicing by creating additional splice sites[69]. It is not clear if these novel splice variants are functional.

More specific A-I editing has been proposed to result in altered protein sequence, through direct codon alteration or splicing alteration, as a method to generate and regulate a variety of different protein isoforms[12]. Most of the known specific edited sites are in genes involved in neurological or synaptic processes[1,34]. In mammals, many of these are in glutamate receptor genes. It is not clear why this bias exists, but could have something to do with the rapid responses required within nerves and at synapses or the differences in the immune system in the brain[1]. An analysis of the abundance of inosine in rat tissues agrees with these observations as the rat brain shows the highest abundance, with one inosine per 17,000 ribonucleotides[110].

1.2.7 A-I RNA Editing & Splicing

The observation that most of the known ECSs occur in adjacent introns shows that A-I RNA editing of these sites must occur before splicing. This section introduces the numerous suggestions in the literature that the processes of A-I editing and splicing are linked. The most obvious example is the self-editing of the rat *ADAR2* transcripts, which results in the creation of an alternative splice site and non-functional protein[22]. The splicing of the *PTPN6* transcript is also affected by editing. The branch site of intron 3 is destroyed through editing of the essential adenosine. This results in the inclusion of the intron, which appears to be involved in the aetiology of acute myeloid leukaemia[111]. This editing site is created by base-pairing of inverted Alu repeats[2].

The *GluR-B* R/G site has been used to model the interaction of editing and splicing[112]. *In vitro* studies show that the splicing machinery and ADAR2 compete for binding to the ECD, and that prior exposure to RNA helicase A removes the ability of ADAR2 to compete. It is possible that the helicase disrupts the hairpin, allowing splicing factors to bind to the otherwise sequestered 5 prime splice site. Given that many ECSs are in introns, the editing system may use hairpins to effectively stall the splicing machinery until editing is complete. An *in vivo* model has been suggested where the CTD (Carboxyl-Terminal Domain) of RNA polymerase II coordinates the editing and splicing, in that order. This may be mediated by *RNA helicase A*, which unwinds the structure once editing is complete[112]. A similar situation can be observed between the *Drosophila melanogaster para* and *mle* genes. Mutations in the *mle* (maleless) gene result in aberrant editing and splicing of the *para* gene[113]. *Mle* is the orthologue of *RNA helicase A* and has been shown to unwind dsRNA structures *in vitro*[114]. Consistent with these theories, both ADAR1 and ADAR2 have been shown to associate with elements of the splicing machinery, including Sm and SR proteins within large nuclear ribonucleoprotein (lnRNP) particles[115].

The *GluR-B* Q/R and the *5HT$_{2C}$R* sites have been shown to have a correlation between the proportion of editing at these sites and the splicing of their adjacent introns. In the *ADAR2* homozygous null mouse, whenever the *GluR-B* Q/R site remained unedited, the next intron was generally retained[35], which suggests that editing is required for efficient splicing. Introducing A-G base substitutions to mimic the A-I substitutions in the *5HT$_{2C}$R* receptor profoundly affected splicing of the normal and upstream splice sites[116], even though these sites occur in the middle of the exon. More recently it was suggested that alternative splicing upstream of edited sites might not be correlated to editing efficiencies, while alternative splicing downstream can be correlated. This was illustrated using two known examples of editing (*Ca α-1T*, a calcium channel & *nAchRα-34E*, a nicotinic acetylcholine receptor)[117].

The regulation of splicing through hairpins and other RNA secondary structures does not appear to be limited to editing sites however. It has been established that alternatively spliced introns are enriched for simple inverted repeats, suggesting that they are somehow involved in the splicing process. While only 44% of the 18 173 genes in the Human Alternative Splicing Database were known to be alternatively spliced, they contained 84% of the 694 237 intronic complementary repeat pairs

(based on criteria described in the study)[118]. Exon skipping is one form of alternative splicing. It has been shown that skipped exons often contain a complementary pair of C-rich and G-rich motifs, which tend to be conserved between mouse and human. It appears that the putative duplexes formed by these sequences may act as the signal to skip these exons[119]. The *Drosophila melanogaster DsCam* gene provides another example of how conserved hairpins might regulate splicing[120]. The gene contains 48 copies of exon 6, only one of which is included in each transcript. The remarkable feat is mediated by a conserved element 5 prime of this group of exons, which is able to form a duplex with complementary 'selector' elements located before each exon 6 variant. This duplex, which can only form with one selector element at a time, relieves splicing repression for the exon, resulting in a transcript containing only that exon 6 variant. Additional examples of secondary structure affecting splicing can be observed for the human *Tau* gene[121] and the mouse *hnRNP A1* gene[122]. It is possible that the ECDs for the known edited sites originally appeared as repeats which were maintained as they regulated splicing, but were then co-opted by the editing machinery. RNA hairpins have also been shown to regulate cellular localisation of transcripts in *Drosophila* oocytes[123].

These observations combine to make a strong case for the interaction of editing and splicing, although a lot more work is required to accurately characterise this interaction.

1.2.8 A-I Editing & Disease

The fact that many of the known mammalian edited genes are implicated in disease processes has lead to a large number of papers detailing the correlation between editing efficiencies of these sites and their associated diseases. Mostly, the correlations seen are negative, or weak. At best these papers show that there is a correlation with editing that may simply be a result of the disease, rather than a causal effect. Table 1.3 (p23) shows a list of these publications and briefly describes the conclusions of each.

Table 1.3. RNA Editing and Disease

Disease/ Result	References	Gene	Association
Epilepsy	124,125	*GluR5/6*	Increased levels of editing in affected temporal cortexes.
Schizophrenia	126	*5HT$_{2C}$R*	Reduced editing efficiency in patients
Major Depression	127	*5HT$_{2C}$R*	Altered editing efficiency in patients
Suicidal depression	127	*5HT$_{2C}$R*	Altered editing efficiency in patients
Suicidal depression	25	*5HT$_{2C}$R*	C site editing is increased, D site editing is decreased. Prozac rescues these changes.
Huntingdon disease	128	*GluR-B*	Putative altered editing efficiency in patients.
Alzheimer disease	128	*GluR-B*	Putative altered editing efficiency in patients.
Malignant Gliomas	129	*GluR-B*	Substantially reduced editing in gliomas. Explains the occurrence of epileptic seizures in association with malignant gliomas.
Sub-acute sclerosing panencephalitis	130	-	Hyper-editing of viral transcripts has been implicated in several deaths by this disease.
Amyotrophic Lateral Sclerosis	39	*GluR-B*	Reduced editing of the *GluR-B* Q/R site correlates with motor neuron death and ALS.
Dyschromatosis Symmetrica Hereditaria	38	Unknown	Mutations in *ADAR1* are implicated in this skin pigmentation disease.

1.3 Previous Methods for Finding A-I Editing Sites

The primary aims of the work in this thesis are to increase the number of known edited sites and to further characterise those sites that have already been identified. Figure 1.4 (p12) shows all the currently known edited sites that result in protein recoding.

The earliest A-I edited sites to be identified were discovered through serendipity. Other sites were identified through direct homology with other known edited sites, such as the glutamate receptor family of edited sites. More recently, however, several groups have made concerted efforts to add to the list of known edited sites. Broadly, these efforts can be divided into biochemical techniques, repeat-based screens, mismatch-based screens, and sequence conservation-based screens.

1.3.1 Biochemical Screens For A-I Edited Sites

Bass *et al* have developed a lab-based technique to identify transcripts containing inosine nucleotides[2,4,110,131]. RNAseT1 cleaves these transcripts *in vitro* at the inosine position, and then the 3 prime end is polyadenylated. Poly-T primers are then used together with an arbitrary primer to amplify these fragments. The PCR products for each arbitrary primer are then run on a sequencing gel and compared to non-RNAseT1-treated products. Any bands observed in the treated gel lane, but not in the untreated lane represent inosine cleavage sites. Using this method they identified 10 nematode edited sites and 5 human sites[2]. All of these sites were found in UTRs, introns or non-coding sequence. The authors suggested that this may reflect an overall bias towards editing in non-coding sequences and that there were probably many more undiscovered sites. One of these sites was in a miRNA and there have been other reports of editing of miRNAs[2,69,132].

Xia *et al* used a different approach to look for *Drosophila* edited sites[133]. They used a polyclonal antibody against inosine to enrich inosine-containing mRNAs from total mRNAs of wild type and *dADAR* mutant flies, respectively. The enriched mRNA portion was amplified and hybridised with *Drosophila melanogaster* cDNA arrays, which identified over 500 potential mRNA edited targets. Sixty-two of these genes also had A-G mismatches observable between publicly available expressed and

genomic sequences. Twelve of these were selected of which seven were experimentally validated. Ohman *et al* are attempting a similar experimental approach (not currently published), except that their antibody is raised against ADAR2 instead of inosine (information based on a poster presented by Marie Ohman at the 2005 Gordon Research Conference in RNA Editing).

1.3.2 Mismatch-Based Screens for A-I Edited Sites

The simplest method to identify novel A-I edited sites relies on a comparison of expressed and genomic sequences from the same regions of the genome. Reverse transcriptase reads inosines as guanosines. Hence, if all the genomic sequences consistently have an adenosine in a particular position, and the expressed sequences have one or more guanosines recorded at the same position, then this is a putative edited site. There are a number of problems with this approach though, including the frequent lack of a sufficient number of sequences, sequencing errors, single nucleotide polymorphisms and mis-alignment issues. These sources of false positives are discussed further in Chapter 3.

This basic concept was used to look for clusters of A-G mismatches between cDNA collections and the genome for both human[134] and *Drosophila*[135]. Stapleton *et al* used the Berkeley *Drosophila* cDNA collection to predict over 30 putative edited sites, although only a few had strong support in the mismatch data. One of these putative edited sites, in an amine receptor gene, was experimentally confirmed. Kikuno *et al* used the Kazusa brain-enriched cDNA collection to predict human edited sites, however the precise locations of these predicted sites were never made available. Both of these groups found that the majority of the putative editing sites they observed were in inverted Alu repeats, which agrees with previous predictions that inosines are predominantly found in non-coding sequences[2]. Mismatch data was also used to help confirm candidates in the inosine pull-down described in the previous section[133]. An unavoidable bias in screens based on mismatches is that they will tend to find sites that are either edited at a high frequency or highly expressed.

Given the apparent bias towards editing in non-coding sequences, it seemed important to specifically target putative edited sites in coding sequences. More complex methods to rank these putative edited sites were also required. Chapter 3 describes the protocol that we developed with these aims in mind. This protocol used

a series of simple filters to remove the majority of the SNPs, sequencing errors and alignment errors. The remaining candidates were analysed for the features of known recoding edited sites, which include high sequence conservation, an ECD structure overlapping the mismatch and clustering of A-G mismatches. Mismatches that resulted in non-synonymous changes or those that were identified in both mouse and human orthologous positions were also recorded. A statistical method was devised to combine each of these features and rank all the putative edited candidate sites. This is the first method to successfully identify more than one of the known mammalian recoding edited sites in a genome-wide screen. A novel candidate, *BC10*, was also identified and validated to be a novel A-I edited target[79].

Levanon *et al* used a similar approach based on a similar set of features. They required the edited region to be conserved above 85% nucleotide identity over at least 50bp and the proposed edit had to be non-synonymous. All remaining mismatches were then screened using a probabilistic algorithm that calculated the likelihood of each mismatch being due to editing. This resulted in two lists of high quality candidates for mouse and human. In addition to several of the known editing sites, they identified four novel edited sites that are found in both lists. These genes were *FLNA, CYFIP2, IGPFB7,* and *BC10*, which was also identified in our screen[81]. In this publication, BC10 was identified under a different name, *BLCAP*. With the exception of *IGPFB7*, these sites have all been experimentally validated.

These analyses demonstrate that searching for A-G mismatches that can be observed in more than one species is a powerful approach. This is because the likelihood of SNPs being conserved between mouse and human is low[136]. In contrast to these two studies, several groups have concentrated their efforts on identifying and characterising the edited sites in non-coding RNA and in repeats.

1.3.3 Repeat-Based Analyses of A-I Edited Sites

In agreement with previous observations, several recent analyses show that the vast majority of human A-I edited sites are situated in introns and UTRs[68-70,134,137]. These analyses all used the alignment of expressed sequences to the human genome. Levanon *et al* specifically looked at A-G mismatches in long inverted repeats and identified more that 12,723 putatively edited sites in 1,637 genes, with an estimated accuracy of 95%[70]. Interestingly, they also found that many of these sites are

erroneously included in the major SNP repository (dbSNP)[138]. Similar results were obtained in other studies, which is not surprising given the similarity of each approach[68,69]. These studies consistently show that ~90% of these novel edited sites are located in Alu repeat sequences[68]. These repeats are found extensively in primate genomes and are not found elsewhere. The mouse has some Alu-like repeats, but they are substantially less common. Kim *et al* showed that the mouse had only 91 cDNAs with significant evidence of editing compared with 2,674 in human. These data suggest that the editing of repeats is exaggerated in primates[68]. A likely cause of this observation is that no other model organism's genome contains any single type of repeat at such a high frequency[71]. In line with this theory, the proportion of editing in Alu repeats tends to be greater when the two inverted repeats share high sequence similarity and are physically close together[69]. It is striking that, in contrast to our efforts, none of these analyses identified any of the known recoding edited sites, however, Athanasiadis *et al* have provided an possible explanation for this. They show that 85% of all pre-mRNAs contain edited Alu repeats. As a result, the number of edited sites in repeats overwhelmed the small number of protein recoding sites. They also show that editing of Alu repeats can result in their exonisation and incorporation into coding sequences[69,134].

Given that the majority of these sites occur in Alu repeat sequences, they cannot be conserved beyond the primate lineage. It is not even clear what function editing of Alu repeats could have. To date, none of the individual sites have been shown to be functional or have a phenotype.

1.3.4 Sequence Conservation Based Screens for A-I Edited Sites

High sequence conservation of the ECD structure appears to be a common feature for the known recoding edited sites. In the fruit fly the levels of sequence conservation can be considerably higher than in vertebrates. For example, the *Drosophila para* gene is exceptionally well conserved across its three edited sites and their ECSs[139]. This observation led Hoopengardner *et al* to screen a collection of over 900 genes involved in neurological functions for high levels of sequence conservation. They visually screened these genes for regions that were highly conserved over 50bp or more between *Drosophila melanogaster* and *Drosophila pseudobscura*. This identified most of the known edited sites in *Drosophila*. In addition, they identified and experimentally validated 16 novel edited sites. All of these genes are involved in

rapid electrical and chemical neurotransmission, and many of the edited sites recode conserved and functionally important amino acids[140]. In a study of ultra-conserved elements in insect genomes, another group identified a number of the known edited sites, although that was not their primary aim[88].

Chapter 4 describes a protocol that uses a combination of sequence conservation and secondary structure to identify novel vertebrate targets and to identify the ECSs for known edited sites. Chapter 5 repeats this analysis for the fruit fly. This work has not been published yet.

In summary, A-I editing has proven to be a widely used mechanism for introducing a number of functional and potentially non-functional edited sites into RNA transcripts. This is a field that has been heavily studied over the last decade, but many questions remain unanswered.

1.4 Bioinformatics, Genomics & Transcriptomics

The work in this thesis relies heavily on the established materials and methods encompassed by bioinformatics, genomics and transcriptomics. A full description of these fields is beyond the scope of this thesis, however, I have attempted to provide a brief overview of the important points.

1.4.1 Bioinformatics & Genomics

Bioinformatics can be described as the collection, organisation and analysis of large amounts of biological data, using computers and databases. This broad definition covers many diverse areas, most of which are not relevant to this thesis. However, the generation and annotation of genome assemblies, and the applications/algorithms designed to analyse this information, have been vital to this work. Much of this work also falls into the field of genomics, which can be defined as the study of genes and genomes.

The human genome project generated vast amounts of genomic sequence data in the form of whole-genome-shotgun reads (WGS) and high-throughput-genomic sequencing reads[141]. Vast amounts of genomic sequences have now been generated for a number of other model organisms. Turning these sequences into the browsable, annotated genomes that are now publicly available has required a wide range of bioinformatic input, including genome assembly programs, DNA alignment programs, protein alignment programs, gene prediction programs, protein secondary structure prediction programs, and protein domain prediction programs[142]. As of October 2005, the Ensembl web browser (www.ensembl.org) held genomes and gene annotation for 21 model organisms[142]. The usefulness of these genomic sequences is not limited to creating assemblies, however. For example, they have also been used to identify SNPs[143].

1.4.2 Transcriptomics

Transcriptomics can be defined as the study of all expressed sequences within a genome. There are many technologies that have been applied to this field, including SAGE[144], CAGE[145], and EST/cDNA sequencing[146] among others. For the purposes

of this thesis, the last two technologies were the most important. Firstly, these gene sequences are extremely useful for guiding gene prediction methods[147]. Secondly, these sequences can be used to observe variations in the transcripts that they are derived from. Among other things, this feature has been used to identify alternative splicing events[148], SNPs[143] and RNA editing sites[79,133-135,138]. Appendix 3 describes my contribution to a Science paper based on the analysis of the mouse transcriptome[73].

1.5 Summary

A-I RNA editing is a process that is linked with a variety of medical conditions. Although there have been several attempts to identify all the edited sites in mammalian and other genomes, it is likely that other protein recoding sites remain to be identified. As of 3 years ago, when this thesis was started, there were no bioinformatic approaches applied to this field. The work in this thesis was intended to fill this niche and identify novel A-I RNA editing sites, both in mammals and other species. This work also aimed to further characterise the known edited sites through bioinformatic approaches. To some extent, both these aims have been successful.

2 Materials & Methods

During the course of this work I have used many techniques and resources more than once. These are divided into materials and methods. The materials used in this work have been divided into computing resources, databases and programs and are discussed below. They are followed by a description of the methods, or multiply used protocols, that I have employed.

2.1 Materials: Computing Resources

Almost all the methods described here have been implemented through Perl scripts run in a Solaris UNIX environment. The remainder were carried out manually. Perl has become one of the most popular languages for biological data analysis for a number of reasons. Primarily it is easy to learn and often has multiple methods for achieving the same task. It also has a highly developed capacity for detecting patterns in text data. Given the enormous amounts of publicly available biological data, and the assumption that there are patterns underlying such data, Perl was an ideal choice of programming language to use.

2.2 Materials: Databases

2.2.1 Ensembl

The Ensembl databases[142] are produced and maintained in a joint project between EMBL-EBI (http://www.ebi.ac.uk/) and the Sanger Institute (http://www.sanger.ac.uk/). During the course of this work I have used several versions of the Ensembl databases between May 2003 and June 2005.

The databases, which can be found at http://www.ensembl.org/, provide a genome-based annotation framework that is consistent between species. This allows for easy comparison of data between organisms. The species covered by these databases include *Homo sapiens* (human), *Pan troglodytes* (chimp), *Mus musculus* (mouse), *Rattus Novegicus* (rat), *Canis familiaris* (dog), *Gallus gallus* (chicken), *Fugu rubripes* (Pufferfish), *Danio rerio* (Zebrafish), *Anopheles gambiae* (Mosquito), *Drosophila melanogaster* (Fruit fly), *Apis mellifera* (Honey bee) and several others.

This list represents a wide variety of biologically interesting and well-studied species, generally with good genome coverage.

The data for each species includes gene and exon predictions, homologous and orthologous gene predictions, coding sequence predictions, splice variants, EST/cDNA alignments and many other features.

Genome sequence files were obtained in FASTA format from the Ensembl ftp site for all the species named above (http://www.ensembl.org/Download/). Exon prediction coordinates and FASTA sequences were obtained from EnsemblMart, which is the web-interface of the Ensembl SQL database (http://www.ensembl.org/Multi/martview).

2.2.2 FlyBase

Although Ensembl contains data for *Drosophila melanogaster*, it does not contain any data for additional fruit fly species. We chose to use *Drosophila pseudoobscura* as a second fruit fly species. The genomic sequences were obtained from FlyBase (ftp://flybase.net/genomes/Drosophila_pseudoobscura/dpse_r10_20041108/fasta/dpse-all-chromosome-r1.03.fasta.gz *and* /dpse-all-scaffold-r1.03.fasta.gz)[149]. This refers to the euchromatin assembly version R1.03, dated 8/11/2005. Annotation files in gff format were also obtained from FlyBase for the same assembly version (ftp://flybase.net/genomes/Drosophila_pseudoobscura/dpse_r10_20041108/gff/*).

FlyBase is an online database of genetic and molecular data for *Drosophila* species. It includes data on a large number of species from the *Drosophilidae* family, and is produced by a consortium consisting of researchers funded by the National Institutes of Health (NIH) and the Medical Research Council (MRC). This database contains genomes, gene and exon predictions, phenotypes, ESTs, cDNAs and many other forms of data. In many ways FlyBase is the *Drosophila* equivalent of Ensembl.

2.2.3 GenBank

GenBank® is the National Institute of Health (NIH) genetic sequence database. As of February, 2004, this contained almost 38,000 million bases in over 30 million

sequence records[150]. These records include both genomic DNA and RNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which also includes the DataBank of Japan (DDJB) and the European Molecular Biology Laboratory (EMBL).

In preparation for the protocol described in Chapter 3, sequence files for all human and mouse expressed sequences were required. Expressed sequence tags (ESTs) were obtained from dbEST (see next section). The remaining expressed sequences were typically full-length cDNAs. These sequences were obtained from the Entrez website (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide), with the following query 'homo sapiens [ORGN]) AND (cdna [TITL] OR mrna [TITL]) AND 100:5000000 [SLEN]' and with the search limits set to 'exclude all the above' & 'molecule = mRNA'. This obtained all sequences with cDNA or mRNA in their title that were between 100bp and 5Mb and were not annotated as ESTs. Mouse data was obtained by substituting 'mus musculus' instead of 'homo sapiens'. These queries resulted in 144,780 mouse cDNAs and 152,223 human cDNAs.

2.2.4 dbEST

dbEST is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences, called Expressed Sequence Tags (ESTs), from a number of organisms[151]. As these ESTs are derived from expressed sequences, they can be used to identify the transcribed regions of genes, predict splice sites and identify expressed sequence variations. ESTs are essentially fragments of larger full-length cDNAs. cDNAs are generally of higher sequencing quality than single pass ESTs. However, ESTs are considerably cheaper to generate than full cDNA sequences. As of June 2005, there were approximately 6 million human ESTs and 4 million mouse ESTs. A large number of additional species also had ESTs in this database, ranging from 900,000 to 1 per species.

All mouse and human ESTs were obtained from GenBank using the command '*homo sapiens* [ORGN] AND gbdiv_est [PROP]' to this URL 'http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucleotide', then selecting a FASTA output format. No limits were required. Mouse data was obtained by substituting '*mus musculus*' instead of '*homo sapiens*'.

2.2.5 Additional Genomic Sequences

The genome assemblies for mouse and human have been generated from overlapping genomic sequences. The human assembly is primarily based on high throughput BAC sequences[141]. The mouse assembly is based on both whole genome shotgun and high throughput BAC clone sequences[152]. The original human sequences were obtained from the EMBL HTG (High Throughput Genomic) repository, while the mouse sequences were obtained from both the EMBL HTG repository and the Ensembl Trace repository.

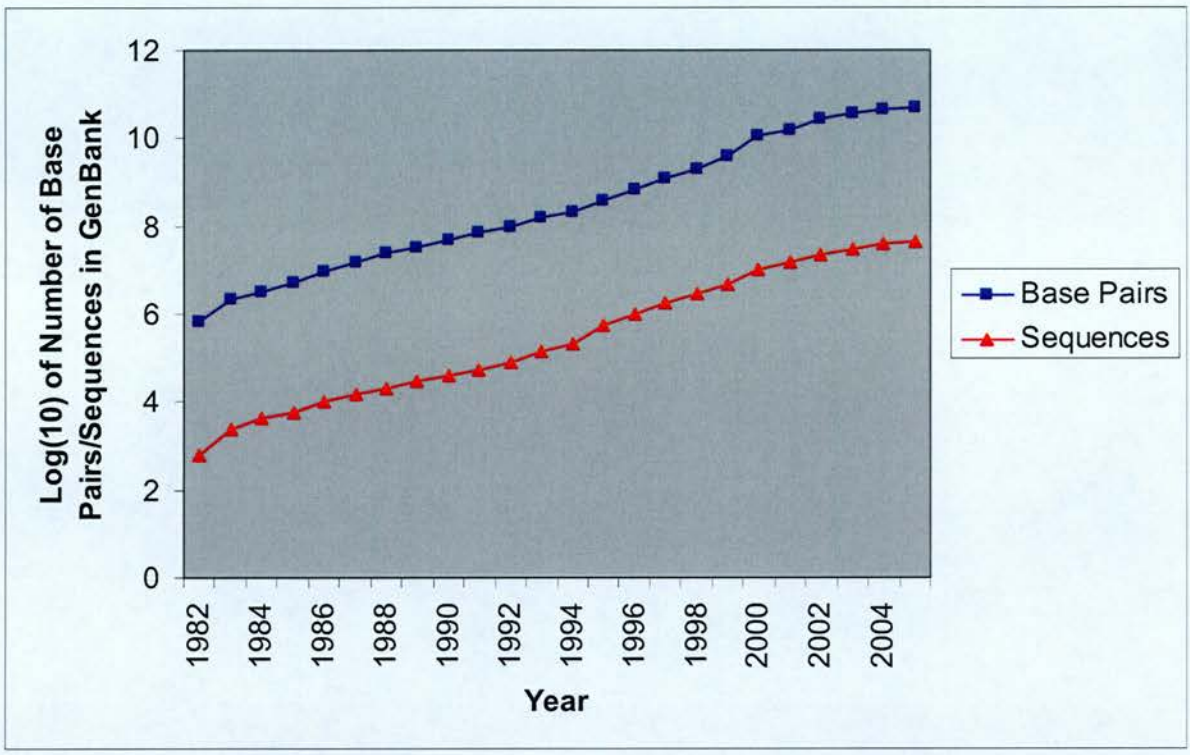
2.3 Materials: Programs & Algorithms

2.3.1 BLAST

BLAST, which stands for Basic Local Alignment Search Tool, is a highly efficient method for rapid searching of nucleotide or protein databases[153]. The BLAST suite of programs is a set of sequence comparison algorithms that were introduced in 1990. At the time they represented a very large reduction in the time taken to search sequence databases, while maintaining high levels of sensitivity. In the last two decades vast amounts of sequences have been generated, leading to significantly larger sequence databases. For example, Figure 2.1 (p35) shows the exponential growth of the GenBank database since 1980. Due to this growth and the BLAST suite's high efficiency and sensitivity, it is undoubtedly the most widely used bioinformatics application available to date.

The secret to the BLAST suite's success is that the programs break the query sequence and database sequences into short fragments of a defined length (W). The optimal length of these fragments or words varies between the types of search being used. In the initial stages of a BLAST search all the words in the query are screened against the words in the databases. The score of the match between a pair of words is based on a substitution matrix, which provides a score for each nucleotide/amino acid in the query word versus each nucleotide/amino acid in the database word. A variety of matrices can be used. Any word matches scoring above a defined threshold score (T) are termed High-Scoring Segment Pairs (HSPs). These HSPs are extended in either direction in an attempt to generate a maximal alignment with a score

Figure 2.1. The Growth of the GenBank Sequence Database.



This figure shows the growth of the entire GenBank nucleotide sequence database since 1982. The growth has been consistently exponential, both in terms of the number of sequences and the number of nucleotides contained in the database.

exceeding a second threshold (S). These alignments are termed Maximal-Scoring Segment Pairs (MSPs).

The BLAST programs also allow for gaps within the alignments. The inclusion of a gap results in a lower alignment score due to a score penalty. An additional penalty is given for extending a gap, although this is normally less than the gap-opening penalty as a single mutational event may result in a gap of more than one residue. These penalties can be modified as required. The resulting alignments are given expectation values (E values), which are based on the length and quality of the match and the total size of the database being searched. The E value is a measure of the probability of a given match occurring by chance in a given database. A low E value indicates a good match, also termed a BLAST hit.

Primarily, I have been using BLASTN, the standard nucleotide versus nucleotide BLAST program. The parameters used for each protocol are described in the appropriate Results chapter. For exceptionally large numbers of BLAST searches MegaBLAST was used. This is a version of BLASTN that is optimised for large numbers of queries and very long sequences.

2.3.2 The LAGAN Toolkit

There are four components in the LAGAN toolkit including a pair wise local alignment program (CHAOS), a pair wise global alignment program (LAGAN), a multiple global alignment program (M-LAGAN) and a 'glocal' alignment program (Shuffle-LAGAN) which generates pair wise global alignments but seeks to account for inversions, transpositions and some duplications[154].

The local alignments generated by CHAOS are used as alignment anchors for the three LAGAN programs. LAGAN then uses a Needleman-Wunsch method[155] to connect the anchors together. The use of anchors reduces the search space for the Needleman-Wunsch algorithm, resulting in very fast alignments. M-LAGAN performs progressive pair wise alignments (using LAGAN), guided by a phylogenetic tree, which results in a multiple alignment. The relative efficiency and accuracy of these programs[156] have made them widely used in the bioinformatics community.

2.3.3 Local Alignment Algorithms

Local alignment algorithms are used to identify regions of local similarity between two sequences, which do not necessarily include the entire length of either sequence. Water and Matcher are two local alignment algorithms from the EMBOSS package of bioinformatic programs[157]. LALIGN is another program that is essentially identical to Matcher, except that the input and output options differ between versions[158]. EMBOSS, which stands for the European Molecular Biology Open Software Suite, is a free Open Source software analysis package aimed at the manipulation of molecular biology data. This package integrates a range of currently available packages and tools for sequence analysis, including Matcher and Water.

Water is a Smith-Waterman local alignment algorithm[159]. In contrast to the BLAST programs, which use a heuristic approach to achieve greater speed, Smith-Waterman programs are guaranteed to identify the best scoring match between two sequences. The main reason this program was selected was that the user is able to modify the scoring matrix used to score the alignments. Normally a standard substitution matrix is used with pre-calculated scores for each possible match or mismatch in the alignment. This matrix reflects the observed frequencies of changes between nucleotides or residues during sequence evolution. In order to identify potential RNA duplexes a novel matrix was created, which contained altered DNA scores reflecting the abilities of each RNA nucleotide to base pair with each other RNA nucleotide. Using Water in this way allows us to take into account the ability of uridine to base pair with both adenosine and guanosine. Details of the derivation of this matrix are given in the results section. Suitable gap opening and extension penalties were also determined empirically.

A limitation of Water is that it only reports the best match. This was not an issue for the protocol in Chapter 3, but this made Water unsuitable for the protocols in Chapters 4, 5 & 6. Matcher/LALIGN is very similar to Water; however, you can specify any number of matches to be reported. This allows us to predict multiple duplexes between two sequences. LALIGN was originally part of the FASTA package[158], and reformatted as Matcher. Although I have used both versions in this thesis, their results are completely equivalent.

2.3.4 RepeatMasker

RepeatMasker is a program that screens DNA sequences from a specified species for interspersed and low complexity repeat sequences[160]. The program results in a full report of all repeat sequences identified and a masked version of the original sequence. More than half of the human genome is repeated sequence, which includes interspersed repeats derived from transposable elements and long genomic tandem, palindromic or interspersed duplications[141]. The basic components of the program are a modified Smith-Waterman alignment algorithm, called 'cross_match', and a collection of repeat sequence libraries for a range of species. This program was used to ensure that putative edited sites in repeat sequences were ignored.

2.4 Methods

2.4.1 Comparative Genomics

One of the major advantages of having numerous genome sequences is the ability to observe sequence conservation and variation between species. This is the basis of comparative genomics, which can indicate how likely a particular sequence is to have a conserved function or common origin between two or more species. This is particularly useful for identifying genes and their exon structures.

One aspect of the known editing sites is that they are often very highly conserved between species[85,140]. Most protein coding sequences show high levels of conservation, especially at the first and second positions in each codon. However, even the third codon positions, often called the wobble sites, tend to be highly conserved around edited sites. This is likely to be due to the strict requirement for the edited regions to form duplex structures. Any variation in these structures could potentially result in a change in the editing pattern for that site, especially if it is near the edited bases[77].

For some editing sites, such as the *GluR-B* R/G site, there is only one highly conserved region, which contains both arms of the required duplex[32]. In other situations, such as the *GluR-B* Q/R site, there are two highly conserved regions, each containing one arm of the duplex[23]. The length of these highly conserved regions varies between sites. On several occasions in this thesis, these observations of high sequence conservation have been used to help identify ECSs for known editing sites, or to predict novel edited sites and their ECSs. The specific methods used are discussed in the respective Results chapters.

2.4.2 Orthologue Prediction

In order to carry out any meaningful sequence comparisons across species, it is desirable to identify homologous and preferably orthologous sequences. However, the identification of orthologues is not trivial. Before I discuss the methods used to identify orthologous genes, it is important to have a clear understanding of the definitions for orthologous, homologous and paralogous genes. Homologous genes

are those that share significant levels of similarity across their entire lengths indicating common evolutionary origin, and incorporate both orthologues and paralogues. Orthologous genes exist in separate species, but derive from the same ancestral sequence after a speciation event in the last common ancestor. They typically have the same function in each species. Paralogous genes derive from a duplication event within a species and their functions can diverge after the duplication[161].

For illustration, the mouse *GluR-B* gene, the human *GluR-B* gene and the human *GluR-C* gene are all homologous. The mouse *GluR-B* and the human *GluR-B* genes are orthologous. The human *GluR-B* and the human *GluR-C* genes are paralogous. Indeed, the mouse *GluR-B* and the human *GluR-C* genes are also paralogous as the event that *initially* produced the pair was duplication prior to primate-rodent divergence rather than a speciation event.

There are a number of methods for identifying orthologous genes. These range from a relatively simple reciprocal BLAST method, to full phylogenetic tree construction. The most commonly used in large-scale analyses is the reciprocal BLAST method, in which each of a pair of genes must select the other as the best BLAST hit in the genome. This works for the simplest case where two genes have simply diverged since a speciation event. However, if duplication has occurred in one or both of the lineages, then only one of the genes could be selected as an orthologue. The other gene, although strictly orthologous, is often ignored. In situations like this confusion may arise as to which gene is the original orthologue, especially if the duplication is recent. In these cases synteny can be used to identify the original orthologous gene. The most accurate method would be to generate a full phylogenetic tree from all the homologous genes; however, this was considered too time-consuming for our purposes. We decided to circumvent these issues by using the Ensembl pre-computed putative orthologue predictions.

Ensembl predicts four categories of putative orthologues, each with varying degrees of confidence. The UBRH category contains those orthologue predictions that are unique best reciprocal hits. The MBRH category contains those orthologue predictions that are one of many best (high-scoring) reciprocal hits (using BLAST). The RHS category contains reciprocal hits that are confirmed by conserved synteny. Finally, the DWGA category contains putative orthologues derived from a whole genome alignment. This latter category is the most common source of orthologues

between the chimp and human genomes. Together these predictions are able to identify the majority of orthologues, while keeping a fairly strict control over the number of false orthologue predictions. In any case, it was desired that the orthologue predictions be fairly liberal, to ensure that all possible orthologue pairs were included. The Ensembl predictions were suitable for this purpose.

For the protocol based on *Drosophila* species, there was only Ensembl data for *D.melanogaster*. This meant that I had to generate my own orthologue predictions between this species and *D.pseudoobscura*. A reciprocal BLAST method was used in this case. The exact methods used are given in the appropriate Result chapter.

So far, only the identification of orthologous genes has been discussed. However, in some cases it is required to know which exons are putatively orthologous within a given pair of genes. This is easily achieved by BLAST searching each exon against all the exons in the other putatively orthologous gene. Details are given in each protocol where this was carried out.

2.4.3 Mismatch Scanning

A-I RNA editing is a process that affects pre-mRNAs in the nucleus[10]. The inosine is read as a guanosine by the translational machinery as well as by the reverse transcriptases used in RT-PCR. The vast majority of the expressed sequence data is derived from mature mRNAs, which would include any edited transcripts. When you compare these sequences to the genomic sequence they are transcribed from, A-G mismatches may indicate that a gene is edited at a given position. For many of the known edited genes these A-G mismatches can be observed in the public databases (See Chapter 3).

Unfortunately, there are a number of other causes of A-G mismatches between expressed and genomic sequences. These include single nucleotide polymorphisms (SNPs), sequencing errors and mis-alignments. The quantity of these errors makes it exceptionally difficult to reliably infer editing from mismatch data alone. The use of multiple genomic sequences can help identify common SNPs, and there are various other methods for removing mismatches not attributable to RNA editing. These are discussed in the appropriate Results chapter. However, the observation of A-G

mismatches can be used to corroborate additional sources of evidence for novel edited sites.

2.4.4 Inverted Repeats and Editing Complementary Sequences

Many of the published editing sites form RNA duplexes (ECDs), which have been shown to be required for A-I editing to occur[1]. When the RNA sequence is linear these duplexes appear as inverted repeats. However, there are some differences between canonical inverted repeats and RNA duplexes. One major difference is that RNA can contain G-U base pairing as well as the canonical A-U and C-G base pairings. Another major difference is that the ECDs for the known protein recoding edited sites tend to be imperfect (i.e. they contain non-base-paired nucleotides, bulges and loops).

There are a number of programs that can identify inverted repeats, but I am unaware of any that use RNA base-pairing specificities. The existing programs also tend not to tolerate gaps or imperfections in the inverted repeats either. To overcome these issues, basic local alignment algorithms were obtained that allowed a substitution matrix to be defined by the user and could cope easily with gaps and mismatches (Water & Matcher/LALIGN – see Section 2.3.3). A matrix was then defined based on observations of the known edited sites. Details of this process are given in Chapter 3.

2.4.5 Relative Entropy and LOD Scores

In a screen where two or more independent features have been measured, a justifiable statistical method for combining the results is required. A relative entropy approach, based on Log-of-Odds (LOD) scores, provides a suitable method[162]. This method requires two distributions, one based on the positive controls and one based on the negative controls. For our purposes we made the assumption that specific editing outside of repeats is rare, hence the negative population of transcripts would approximate to the whole population (with the known editing sites removed).

The proportions of the positive distribution and negative distributions with a given score are obtained. The odds ratio is the ratio of these two proportions, which gives

the likelihood of a result with that score being from the positive distribution. The log2 of the odds ratio is the LOD score. The main advantage of the LOD score is that, as long as each feature is independent, LOD scores from different features are additive. In practice the scores are often collapsed into bins, due to limited numbers of positive controls.

2.5 Miscellaneous

2.5.1 Sgrab – Rapid Sequence Retrieval System

Sgrab is a Perl sub-routine originally written by Martin Taylor, and modified for use in this thesis. This program allows sub-sequences to be retrieved from large sequence files, such as genome or chromosome sequence files, without having to read in the whole file. This relies on an index, which records the byte index of each FASTA sequence within the file and the length of the FASTA lines following the header. From this any sub-sequence of any sequence in the file can be retrieved more rapidly than with fastacmd, the equivalent utility from the NCBI BLAST package.

3 Results: Mismatch-Based Screen for A-I Editing

3.1 Preface

Recent studies have demonstrated widespread adenosine-inosine RNA editing in non-coding sequence, however the extent of editing in coding sequences has remained unknown[70]. For many of the known sites, editing can be observed in multiple species and often occurs in well-conserved sequences[1]. In addition, Figure 1.4 (p12) shows that they often occur within imperfect inverted repeats and in clusters. Here we present a bioinformatic approach to identify novel sites based on these shared features. Mismatches between genomic and expressed sequences were filtered to remove the main sources of false positives, and then prioritised based on these features. This protocol is tailored to identifying specific recoding editing sites, rather than sites in non-coding repeat sequences.

The protocol described in this chapter is more sensitive for identifying known coding editing sites than any previously published mammalian screen. A novel multiply edited transcript, *BC10*, was identified and experimentally verified. *BC10* is highly conserved across a range of metazoa and has been implicated in two forms of cancer. This majority of the work in this chapter was published in Bioinformatics (Clutterbuck *et al*, 2005 Bioinformatics 21:11, 2590-5).

3.2 Introduction

A number of analyses have tried to identify A-I edited sites through alignment of expressed and genomic sequences[68-70,81,134,135,137,163]. The most recent of these screens identified several of the known edited sites and an additional four edited sites. However, before this publication none of the known edited sites had been identified by any of these screens and most only predicted sites that result from inverted pairs of high copy number repeats. Only one non-repeat mediated protein recoding site has been identified in any of these screens and this was in the fruit fly[135].

The main problem with these approaches is that the A-G mismatches from specific protein recoding edits are diluted by vast numbers of A-G mismatches from repeats, single nucleotide polymorphisms, sequencing errors and mis-alignment errors. However, there are several features that can be used to help distinguish the genuine protein coding edits from the repeat-mediated sites and the false positives.

It has been established that many of the known recoding A-I editing sites occur in clusters, are well conserved and are found in imperfect inverted repeats, which form RNA duplexes, which are termed ECDs in this thesis[23]. This is illustrated in Figure 1.4 (p12) and Table 3.1 (p46). These features have been reviewed and analysed to determine their predictive power for identifying novel RNA editing sites. We have also attempted (unsuccessfully) to identify other features common to these sites including similarities in local secondary structures or motifs.

A combination of seven predictive features have been used to screen a large set of expressed versus genomic sequence mismatches, including suitable filters to remove SNPs, sequencing errors and alignment errors. In contrast to previous screens this method successfully identified many of the known A-I recoding sites as well as identifying a novel experimentally validated recoding site.

Table 3.1. Sequence Conservation and ECS Predictions for the Known Recoding Edited Regions

Edit Type	Edited Gene	Protein Modification	Sites in Cluster	Seq. Cons.[a] over 120bp(%)	ECD Score[b]	Relative ECS Position[c]	Agrees with Known ECS Location[d]	Identified by our Protocol[e]
A-G	*GluR-B*	Q/R	2	99.2	112	+324	N	Yes – Ranked 3rd highest scoring gene.
A-G	*GluR-B*	R/G	1	92.5	106	+68	Y	Yes – Ranked 3rd highest scoring gene.
A-G	*GluR-5*	Q/R	1	95.8	69	-300	N	Yes - Ranked poorly due to very weak ECS
A-G	*5HT2cR*	Various	4	95.8	121	+143	Y	Yes – Ranked 1st highest scoring gene.
A-G	*GluR-C*	R/G	1	98.3	140	+58	Y	Yes – Ranked 5th highest scoring gene.
A-G	*GluR-D*	R/G	1	95.0	98	+61	Y	Yes – Ranked 6th highest scoring gene.
A-G	*KCNA1*	I/V	1	99.2	76	-1537	-	No – Mismatch not observed
A-G	*ADAR2*	Alt. Splice[f]	8	-	NA	NA	-	No – It is intronic
A-G	*GluR-6*	I/V	2	83.3	72	-1490	-	No – It was not in Mouse Ensembl 30
A-G	*GluR-6*	Q/R	1	81.7	69	+1717	N	No – It was not in Mouse Ensembl 30

a) Sequence conservation is generated as percentage nucleotide identity between 120bp surrounding the mouse site and the orthologous human sequence. b) ECD score is generated by an alignment algorithm using RNA base-pairing specificities (see Materials & Methods). c) Location of the 5'end of the ECS, relative to the editing site (+ indicates downstream). d) If the predicted ECS agrees with the published ECS this is recorded with a 'Y'. e) Brief comments on how well each gene ranked or why they were not found. f) Editing of *ADAR2* results in alternative splicing by creating an alternative 3' splice site.

3.3 Materials and Methods

An outline of the method is given in Figure 3.1 (p48).

3.3.1 Materials: Sequence Data

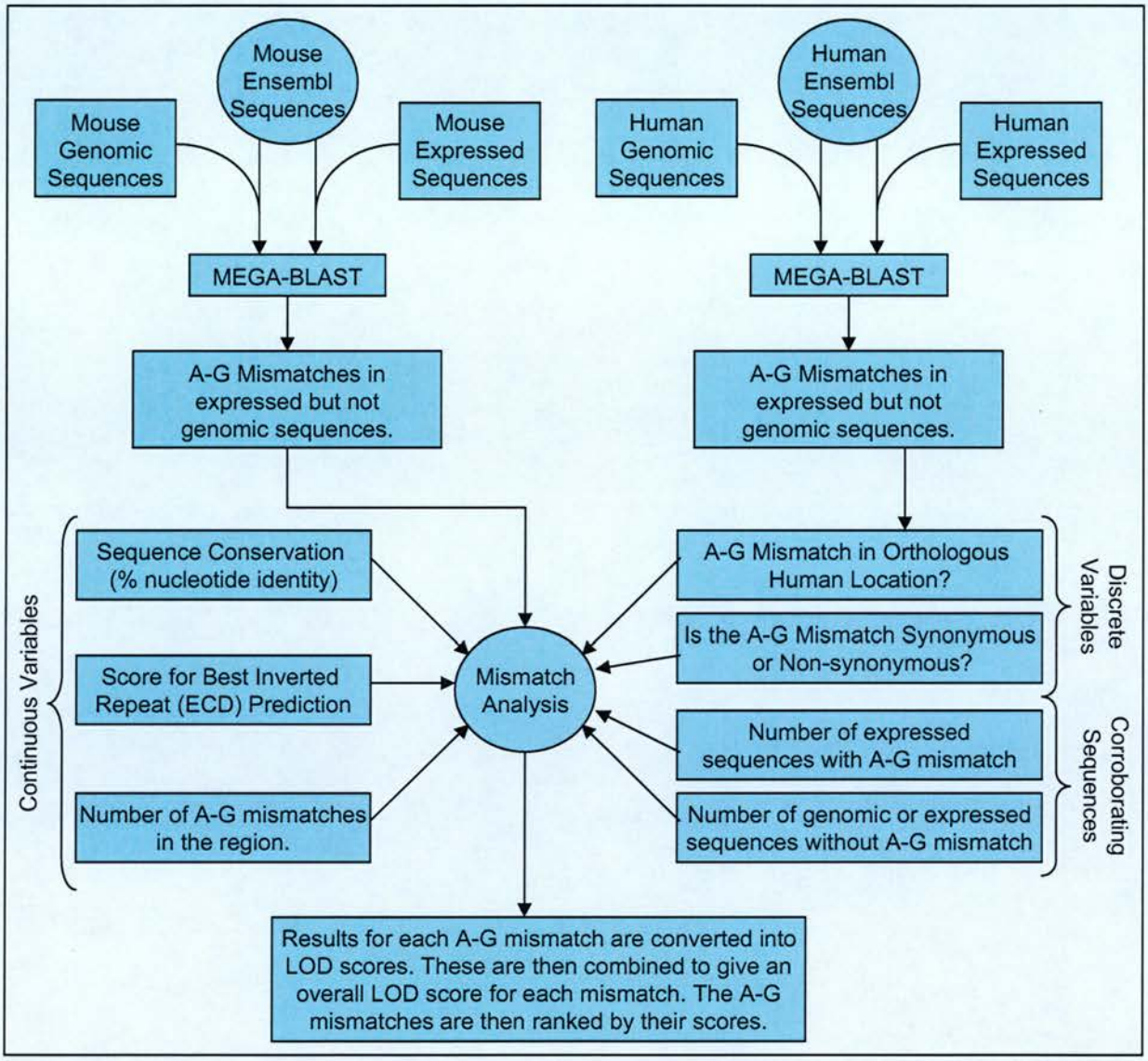
The analysis presented here is based on mouse and human, as large amounts of expressed and genomic sequence data are required. We have focused on the mouse as most of the mouse data is from a single strain, which reduces noise from single nucleotide polymorphisms. Human data are only used to confirm the putative mouse editing sites. To reduce the computational expense of this project to a manageable level, we required a collection of reference sequences that contained non-redundant sequences for the majority of the known genes for each species. For this purpose, concatenated exon sequences were obtained from Ensembl[142] (based on mouse NCBI version 30 and human NCBI version 33). These included most of the known exonic recoding editing sites, with the exception of *GluR-6* which was obtained from GenBank[150] using the accession NM_010348. Orthologous mouse-human pairs were obtained from Ensembl. Where multiple homologues were predicted, we analysed each of them for sequence conservation and putatively orthologous mismatches, and then used the best homologous gene in the remaining analyses.

An editing region is defined here as an exon with one or more editing sites. A necessary limitation of this protocol is that editing sites would not be observed if they are intronic or do not occur within an Ensembl transcript. This was the case for three of the known RNA editing regions (*ADAR2* and *GluR-6* Q/R & I/V sites). The remaining known editing sites constituted the positive control set for this protocol.

3.3.2 Methods: Identifying Mismatches

Mismatches were identified using MegaBlast[153] (version 2.2.6) searches of mouse and human Ensembl genes against all publicly available expressed and genomic sequences for their respective species. BLAST matches that were less than 100bp long, or less than 98% nucleotide identity, were discarded. This threshold removed low quality sequences and matches from homologous genes, but it also removed one

Figure 3.1. An Overview of the Mismatch Based Screen for A-I Edited Sites



All the publicly available expressed and genomic sequences are BLAST searched against a set of Ensembl genes. From these alignments, A-G mismatches that are found in the expressed sequences, but not in the genomic sequences are obtained. These are then analysed for seven features. These are divided into continuous variables, discrete variables and corroborating sequences. LOD scores are obtained for the results of each feature. These are based on the distribution of the known edited sites compared to all other A-G mismatches for a given feature.

edited EST from a known editing site. Unknown edited transcripts may also have been removed. Expressed sequences were obtained from dbEST[151] and GenBank[150]. Genomic sequences were obtained from the EMBL HTG repository for both mouse and human. Mouse shotgun trace repository sequences were obtained from Ensembl[142]. All sequences were up to date as of September 2003. Clone and strain data were obtained from GenBank[150]. Clone identifiers were used to remove redundancy from the set of expressed sequences. An initial set of 28,992 A-G mismatches found between the genomic and expressed sequences, but not between genomic sequences was constructed.

3.3.3 Methods: Analysing the Mismatches for Features of Edited Sites

Each of these A-G mismatches could potentially have been the result of editing. In order to prioritise these mismatches they were analysed for a series of seven features. These are as follows:

A) The number of putatively edited mouse cDNAs or ESTs with the same mismatch at the same position (*Allowed values: 1,2,>2).* This is the number of sequences that support the prediction of an edited site.

B) The number of non-edited mouse cDNAs or ESTs combined with the number of publicly available genomic sequences for each given mismatch (*Allowed values: 1,2,>2).* This is the number of sequences for this position that do not support the prediction of an edited site.

C) Where possible the human homologues were aligned using Lagan[154]. Putative mouse sites were considered to be conserved in human if there were also A-G mismatches in the orthologous/equivalent location in human expressed sequences *(Allowed values: Y,N).* This provided strong evidence that a mismatch was not a SNP or a sequencing error.

D) The effect of the edit on the amino acid sequence was calculated by BLAST[153] searching the Ensembl nucleotide sequence against the equivalent protein sequence, then mapping the putative editing site onto the alignment. This allowed us to distinguish between edits that alter the amino acid sequence and those that do not (*Allowed values: Synonymous, Non-synonymous).* Edits that affected the coding sequence were considered more interesting, and more likely to be functional. Most of the published edited sites in coding sequences are non-synonymous.

E) Sequence conservation was analysed using the same Lagan mouse/human alignments, from which the best conserved 120bp window overlapping each putative editing site was selected (*Continuous variable – percentage identity over 120bp).* Most of the published edited sites are well conserved between mouse and human (see Table 3.1 – p46).

F) Putative mouse ECSs were identified by scanning for inverted repeats using a Smith-Waterman local alignment algorithm from EMBOSS[157], Water, based on a scoring matrix modified for RNA base pairing specificities. Details of this are given in Section 3.4.1. The alignment was generated between a 70bp region flanking the editing site and a reversed flanking 4Kb region, extending 2Kb in either direction from the middle of the 70bp region (Note: the sequence is reversed, not reverse complemented). This test is unavoidably biased against edits that occur towards the end of inverted repeats (*Continuous variable – the local alignment score).* Figure 1.4 (p12) shows that most of the known recoding sites occur in inverted repeats that form RNA duplexes.

G) Clusters of sites were defined by the observation of more than one putative editing site within an exon (*Continuous variable – number of sites in region).* Figure 1.4 (p12) shows that several of the known edited sites occur in clusters.

3.3.4 Combining the Results with LOD Scores

The results of these analyses were combined using a relative entropy approach[162]. For a given feature i with a value x_i, we assigned a log-odds (LOD) score:

$$s_i(x_i) = \log_2\left(\frac{f_i(x_i)}{g_i(x_i)}\right), \qquad S = \sum_{i=1..7} s_i(x_i).$$

where $f_i(x_i)$ was the proportion of all the positive controls in an interval containing feature value x_i and $g_i(x_i)$ was the proportion of the remaining 28,992 A-G mismatches in the same interval. The proportions used were then smoothed to avoid over-fitting to the limited number of positive controls[162].

The overall score assigned to a putative editing site was the sum of the LOD scores for these seven features. This method required discrete distributions for each feature, for both the foreground (positive control regions) and the background (all other A-G

mismatches). Four of the features are already discrete (features A, B, C & D), and the three continuous features were split into 15 discrete bins, each containing a 15th of the background frequencies. To prevent over-fitting of the positive controls we smoothed both the foreground and background distributions for all the features. This was carried out by two iterations of a smoothing filter of width 3 bins, where α is the weight for the central value, β is the weight for the values to the left and right, and Ω is the number of bins.

$$\beta = \frac{\Omega}{100}, \qquad \alpha = 1 - (2 \times \beta).$$

For fifteen bins, this equated to weights of 0.7 and 0.15 respectively. Smoothing the distributions ensured that any candidates that did not exactly match the positive controls could still score well. Next, the ends of the distributions were collapsed, merging all the bins containing no positives with the last bin that contains a non-zero frequency for the positive controls. As a second measure to ensure that we were not over-fitting to the data, we applied a jack-knife (leave-one-out) method. When we scored each of the positive controls, the frequencies of the positive controls in each bin did not include the particular control being scored. In addition we removed all closely related sites. For example, when the *GluR-C* R/G site was being scored, *GluR-B* R/G, *GluR-D* R/G sites and the *GluR-C* R/G site itself, were all removed from the foreground distributions used to calculate the LOD scores.

3.3.5 Dealing with SNPs, Sequencing Errors and Mis-Alignments

The mismatches were ranked by their combined LOD scores. This dataset was expected to contain many single nucleotide polymorphisms, sequencing errors and alignment errors. Potential inter-strain SNPs could be identified where the edited sequences are not from the BL6/C57 strain, which is by far the most common source of sequence data for the mouse. Strain data was available for the majority of expressed sequences[150]. Intra-strain SNPs could not be identified, but should have been rare due to the highly inbred nature of the BL6/C57 strain. Sequencing errors were selected against by the scoring system, which includes measures of the number of edited and unedited sequences. Mismatches with two or more edited sequences were highly prioritised in this way. Alignment errors appeared in two main forms. Firstly, the Ensembl exon predictions were occasionally inaccurate, leading to false mismatches at splice junctions (especially at non-canonical splice sites[164]).

Alignments were generated, using EST2Genome[165], for the top 50 genes and any mismatches that could be explained in this way were discarded. Secondly, homologous sequences sometimes aligned to the wrong Ensembl gene if they were greater than 98% identical over greater than 100bp. To remove these occurrences, all mouse Ensembl genes were BLAST searched against both the genome and each other to identify the positions where close homologues differed. Any mismatch that could be explained by the existence of a homologous sequence that varies at that position was discarded.

As most of the positive controls have matching sequences that are edited in brain (and in the light of the results of Morse *et al*[2] and Hoopengardner *et al*[140]), we annotated any mismatches that had edited sequences from brain or neural tissues. The list of mouse brain/neural tissue derived expressed sequences was obtained from GenBank[150]. RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) was used to identify all the repeats in the Ensembl mouse genes. Any putative sites that were within a repeat were annotated accordingly. None of the top 20 genes had recoding sites overlapping a repeat. PHD was used to predict the protein secondary structure and the solvent exposure of the edited sites[166].

3.3.6 Experimental Validation of the Candidate Edited Sites

The top ten novel candidate sites were experimentally tested in Mary O'Connell's lab (MRC HGU, Edinburgh). This experimental work was carried out by Anne Leroy. RT-PCR was used to confirm the sequences of expressed sequences, while standard PCR was used to confirm the genomic sequences. Whole RNA from C57BL6/J mice heart or brain was isolated using TRIS-REAGENT (Sigma) following the instructions of the supplier. RT was performed using M-MLV Reverse Transcriptase from Promega, 1μg of RNA and 1mM of RT primer. 2nd strand synthesis were performed by PCR using HIFI enzyme (Roche) as recommended by the supplier and 1/10 of the RT reaction as template. Where products of the correct size were obtained, they were sequenced (ABI PRISM BigDye Terminators, MRC sequencing service, Edinburgh, Scotland, UK).

The top four novel candidate edited sites also had individual clones tested. Products from two or more independent PCR reactions were gel purified (Qiagen) and cloned

into the PCR cloning vector pGEMTeasy (Promega). Clones were sequenced on both strands (ABI PRISM BigDye Terminators, MRC sequencing service, Edinburgh, Scotland, UK). Genomic sequences of the genes of interest were checked. DNA from C57BL6/J mice heart or brain was purified and amplified using HIFI enzyme (Roche). The PCR product was gel-purified and sequenced on both strands (ABI PRISM BigDye Terminators, MRC sequencing service, Edinburgh, Scotland, UK). The primers used for brain and heart were identical.

3.4 RESULTS

3.4.1 Making the RNA Matrix

In order to identify potential ECDs with a local alignment algorithm, a suitable matrix had to be constructed. DNA and RNA base-pairing specificities are not the same. DNA only base pairs through the canonical pairs of A-T and G-C. However, in RNA, G can also base pair with U (the equivalent of T). To create an RNA matrix a standard DNA matrix was obtained from EMBOSS[157]. This had scores of +6 for a G-C pair, +5 for an A-T pair and –12 for any other combination. The G-C pair receives a higher score as this interaction results from three hydrogen bonds compared to two for A-T pairs, and is therefore more stable[167]. It was reasonable to assume that RNA base-pairing specificities for these pairs would probably be similar so these scores were kept. However, to allow for G-U base pairing, the score for this interaction was changed from –12 to +3. Other values ranging from +1 to +5 were also tried, but +3 seemed the most successful based on practical experimentation on known RNA duplex structures, including the known edited sites. Although a G-U base-pair is stabilised by two hydrogen bonds, the score of +3 reflects that fact that this is a less thermodynamically preferable interaction[167]. The score of +3 reflects the fact that this interaction has two hydrogen bonds to stabilise it, but it is not a preferable interaction. These same RNA duplex structures were also used to determine useful gap opening and gap extension penalties through a process of trial and error with the known editing sites.

The resulting matrix is as follows:

```
   A   U   G   C   N
A -12   5 -12 -12 -12
U   5 -12   3 -12 -12
G -12   3 -12   6 -12
C -12 -12   6 -12 -12
N -12 -12 -12 -12 -12
```

The use of this matrix in conjunction with a suitable local alignment algorithm provided a very fast way to search for putative RNA duplex structures. In contrast to the currently available algorithms for predicted RNA structure this method can search very large sequences without prohibitive memory and processing requirements.

3.4.2 Analysis of Known Editing Sites

Recoding A-I edits tend to be conserved across related species[168]. Eleven out of the twelve A-I recoding mouse edits shown in Table 3.1 (p46) were supported by A-G mismatches in the public databases, and eight of these were also observed in human expressed sequences (although four of these are from the same cluster in the *5HT$_{2C}$R* gene). The levels of sequence conservation observed between the A-I mouse and human editing sites varied between 99% and 82% identity over 120bp around the editing site (Table 3.1 - p46). This high level of conservation agrees with previous observations[140] and suggests that sequence conservation is a useful predictor of recoding editing sites.

Several of the known recoding sites have published ECSs in nearby introns[23,24,32,63,78,169]. Our novel method for finding mouse ECSs was able to correctly identify four out of seven of these ECSs. Putatively orthologous ECSs were also observed in human for these four ECSs. The remaining ECSs, overlapping the *GluR-B, GluR-5* and *GluR-6* Q/R sites, were not identified due to the identification of higher scoring putative ECSs nearby. We also identified putative ECSs for several of the other positive controls, although they were generally weaker and did not tend to occur in the 3' introns as with most previously characterised ECSs (see Table 3.1 – p46).

One feature of many recoding sites is that in addition to one highly edited adenosine, other nearby adenosines are also edited, although to a lesser extent. We define an editing region as an exon that contains one or more known editing sites. Table 3.1 (p46) demonstrates the usefulness of a cluster analysis as four out of the ten A-I regions contain a cluster. Finally, all of the known A-I editing regions contain at least one recoding edit, many of which have been shown to be functional and some have been implicated in disease.

Through simple BLAST searches of the publicly available cDNA and EST databases, we were able to identify edited sequences expressed in the nervous system for all the mammalian A-I recoding edits, except for the *KCNA1* site and the *ADAR2* site. This observation agrees with previous reports on mammalian[168] and *Drosophila* recoding editing sites[140] that suggest that most A-I recoding edits are specific to the

brain and associated tissues. Although this could be a powerful predictor of novel editing sites, it would introduce bias against editing in other tissues so we have not included it in this analysis.

The remaining known sites were not identified, as the *ADAR2* site is intronic[22], the *GluR-6* edited exons were not included in the Ensembl gene set, and there were no expressed sequences with mismatches to the *KCNA1* site. The *KCNA1* site emphasises the limitations of the available expressed sequence data demonstrating that some sites may be missed. This analysis was not applicable to C-U, U-C, or U-A editing as the known edited sites could not be identified due to a lack of expressed sequence data or their absence from Ensembl genes.

To determine the discriminatory power of each of these features, we compared the proportions of the positives above specified thresholds, to the proportion of all A-G mismatches above these thresholds (see Table 3.2 – p57). Thresholds were selected to include most or all the positive controls, while minimising the proportion of the remaining mismatches above each threshold. These discriminatory ratios demonstrate the power of each of these features, but they were not used further in this analysis. These thresholds are based on the A-I positive control set. The observation of the editing site in both mouse and human is the most powerful predictor for genuine editing sites with a discriminatory ratio of 63. Sequence conservation and the presence of a strong putative ECS are also good predictors (discriminatory ratios of 8.5 and 10.3). Observations of clustering, more than one edited sequence and a coding change, are also useful predictors, although not as powerful as the previous three features. Almost all mismatches had more than one supporting non-edited sequence, which suggests that the observed mismatches are not due to errors in the genomic sequences. The discriminatory ratio for editing in brain/neural tissue is particularly high (29) in agreement with previous observations that all known A-I recoding sites are edited in these tissues[1,140]. However, use of this analysis would bias the results against potential editing in other tissues so we have not used it in the main screen.

Table 3.2. Discriminatory Ratios for A-G Mismatch Features Analysed

Feature Analysed	Threshold[a]	A-I Positive Controls Above Threshold	All A-G Mismatches Above Threshold[b]	Discriminatory Ratio[c]
Edited in Mouse and Human expressed sequences	Yes	50%	0.8%	63x
ECD predictions	>95	55%	5.4%	10.3x
Sequence Conservation	>95%	67%	7.8%	8.5x
Num Mouse Edited Sequences	>1	57%	22%	2.6x
Coding Change	Yes	100%	41%	2.4x
Clustering	>1 site	40%	30%	1.3x
Num Non-Edited Mouse Sequences	>1	100%	98%	1x
Editing in Brain/Neural Tissues[d]	>1 transcript	100%	3.5%	29x

Table 3.2. Proportions of Positive Control Regions versus All A-G Mismatches Above Thresholds for Analysed Features. a) We selected suitable thresholds based on the positive control set. b) This set contains 28,992 mismatches between mouse expressed and genomic sequences. c) The discriminatory ratio is the ratio of the percentage of A-I positive control regions over the threshold versus all A-G mismatches over the threshold. d) Although the observation of editing in brain/neural tissue is a strong predictor for the positive controls, it may introduce bias against potential editing in other tissues. This feature is not used in the final scoring method.

3.4.3 Genome-wide Identification of RNA Editing Sites

To screen all A-G mismatches in the genome, a relative entropy approach[162] was used to combine the results of the seven editing site features (see Materials and Methods). The positive control set consisted of the ten recoding sites (from seven edited regions) that are included in Ensembl transcripts. To combat over-fitting to the positive control set we applied an iterated smoothing operation to the frequency distributions before generating LOD scores. In addition we used a conservative jack-knife approach which ensured that each positive control was scored using only the non-related positive controls (see Materials & Methods).

This was the first mammalian screen to identify any of the known recoding editing sites. Figure 3.2 (p62) shows the distributions for the three continuous variables (sequence conservation, clustering, and ECS score) and the total LOD scores of all A-G mismatches. This figure demonstrates that the three variable features and the total LOD score are useful and efficient for distinguishing the positive controls from the remaining mismatches.

This scoring system identified seven out of the ten positive control edits in the ten top ranked mismatches (including the *GluR-B* Q/R, *GluR-C R/G, GluR-D R/G,* and all four *5HT$_{2C}$R* edits). Of the three remaining positive controls, *KCNA1* was missed as we did not find any matching edited sequences, while the rest were ranked badly due to poor sequence conservation, poor ECS predictions, the lack of clusters or the lack of orthologous mismatches in human.. A list of the 20 top ranked edits is given in Table 3.3 (p60). For the two positive control genes containing clusters of known exonic editing sites, we successfully identified each individual site.

IMPORTANT NOTE

Early in the progression of this project a false positive became mistakenly incorporated into the positive control set. This was an N/S recoding site in the GluR-D gene that corresponded to an N/S change that is observed when the flip and flop exons are compared. This false site was predicted to contain a cluster of sites, which means that the cluster analysis of this protocol was overly biased towards sites with clusters.

In order to account for this, Table 3.3 (p60) also shows the Top 20 ranked sites when the clustering LOD scores are removed (Sections C). Although there are some differences in the rankings, these two Top 20 lists broadly have the same composition. This would have been predicted given the low discriminatory ratio for clustering (i.e. 1.3). In retrospect, although this is very unfortunate, I feel that this disparity falls within the levels of error that can be expected from any protocol of this complexity.

Unless otherwise stated, the following analyses refer to the top 20 ranked edits based on all seven features (including the cluster LOD scores).

3.4.4 Novel Candidate A-I Edited Sites

The 20 top-ranked edits included 13 novel candidate A-I editing sites. The gene descriptions in Table 3.3 (p60) show that these putative edited sites were in a variety of genes. The ten highest scoring novel candidate-recoding edits were experimentally tested for evidence of editing. We tested mouse brain and heart RT-PCR products for evidence of editing at the predicted sites. The brain was chosen as all the known A-I recoding sites are edited in this tissue, and the heart was chosen as a control. Athanasiadis *et al* tested brain and lung for the same reasons[69]. For nine of the top ten novel candidates, there was no evidence of editing. It is possible that these genes are edited in other tissues, however exhaustive testing of these sites in every tissue and developmental stage was beyond the scope of this work. The top ranking novel edit, *BC10*, contains a novel edited region, which we have experimentally verified (Figure 3.3 – p64).

We tested 17 other novel candidate editing sites, in addition to the ten highest scoring novel candidates. These candidates were randomly selected and vary widely in their score ranks. None of these candidates showed experimental evidence of editing, which suggests that recoding editing is only common at the top end of the distribution, and confirms the reliability of the scoring system.

Table 3.3. The Top 20 A-G Mismatches from the Mismatch-Based Screen (Section A)

Rank	Ensembl Mouse 30 Gene ID	LOD Score	In Human	Num. Edited Transc.	Num. Un-Edited Transc.	Num. Genomic Seqs.	Cluster Size	Cons 120 (% identity)	ECD/IR Score	Coding	Coding change	Comment / Experimental Status
1	ENSMUSG00000041380	13.1	Y	2	2	10	4	95.8	121	Y	I/V	Positive Control - 5HT2cR
2	ENSMUSG00000041380	13.1	Y	2	2	10	4	95.8	121	Y	N/S	Positive Control - 5HT2cR
3	ENSMUSG00000047214	12.3	Y	6	37	10	2	99.2	244	Y	Q/R	Tested & Verified - BC10
4	ENSMUSG00000041380	11.0	Y	3	1	10	4	95.8	121	Y	I/M	Positive Control - 5HT2cR
5	ENSMUSG00000041380	11.0	Y	3	1	10	4	95.8	121	Y	I/V	Positive Control - 5HT2cR
6	ENSMUSG00000033981	9.5	Y	8	0	8	2	99.2	112	Y	Q/R	Positive Control - GluR-B Q/R
7	ENSMUSG00000030776	8.9	Y	2	11	44	2	99.2	91	Y	D/G	Tested - No editing observed
8	ENSMUSG00000001986	8.4	Y	1	0	10	1	98.3	140	Y	R/G	Positive Control - GluR-C R/G
9	ENSMUSG00000025892	7.9	N	2	2	37	5	95.0	98	Y	R/G	Positive Control - GluR-D R/G
10	ENSMUSG00000020608	7.6	N	1	5	14	5	94.2	109	Y	K/E	Tested - No editing observed
11	ENSMUSG00000033456	7.5	Y	1	3	5	4	95.8	81	Y	K/E	Tested - No editing observed
12	ENSMUSG00000026014	7.4	N	3	0	12	1	95.8	123	Y	D/G	Tested - No editing observed
13	ENSMUSG00000021743	7.4	Y	1	9	26	2	95.0	96	Y	K/E	Tested - No editing observed
14	ENSMUSG00000017776	7.3	Y	2	13	15	1	100.0	69	Y	E/G	Tested - No editing observed
15	ENSMUSG00000030256	6.7	N	1	0	18	4	95.8	75	Y	E/G	Tested - No editing observed
16	ENSMUSG00000026469	6.1	N	3	4	21	3	93.3	100	Y	N/D	Tested - No editing observed
17	ENSMUSG00000049939	5.9	N	2	6	21	5	100.0	68	Y	E/G	Tested - No editing observed
18	ENSMUSG00000050587	5.9	N	1	11	6	9	98.3	73	Y	K/E	Not tested
19	ENSMUSG00000024498	5.9	N	2	0	7	1	96.7	93	Y	K/E	Not tested
20	ENSMUSG00000004044	5.9	N	2	0	16	1	95.8	97	Y	K/E	Not tested

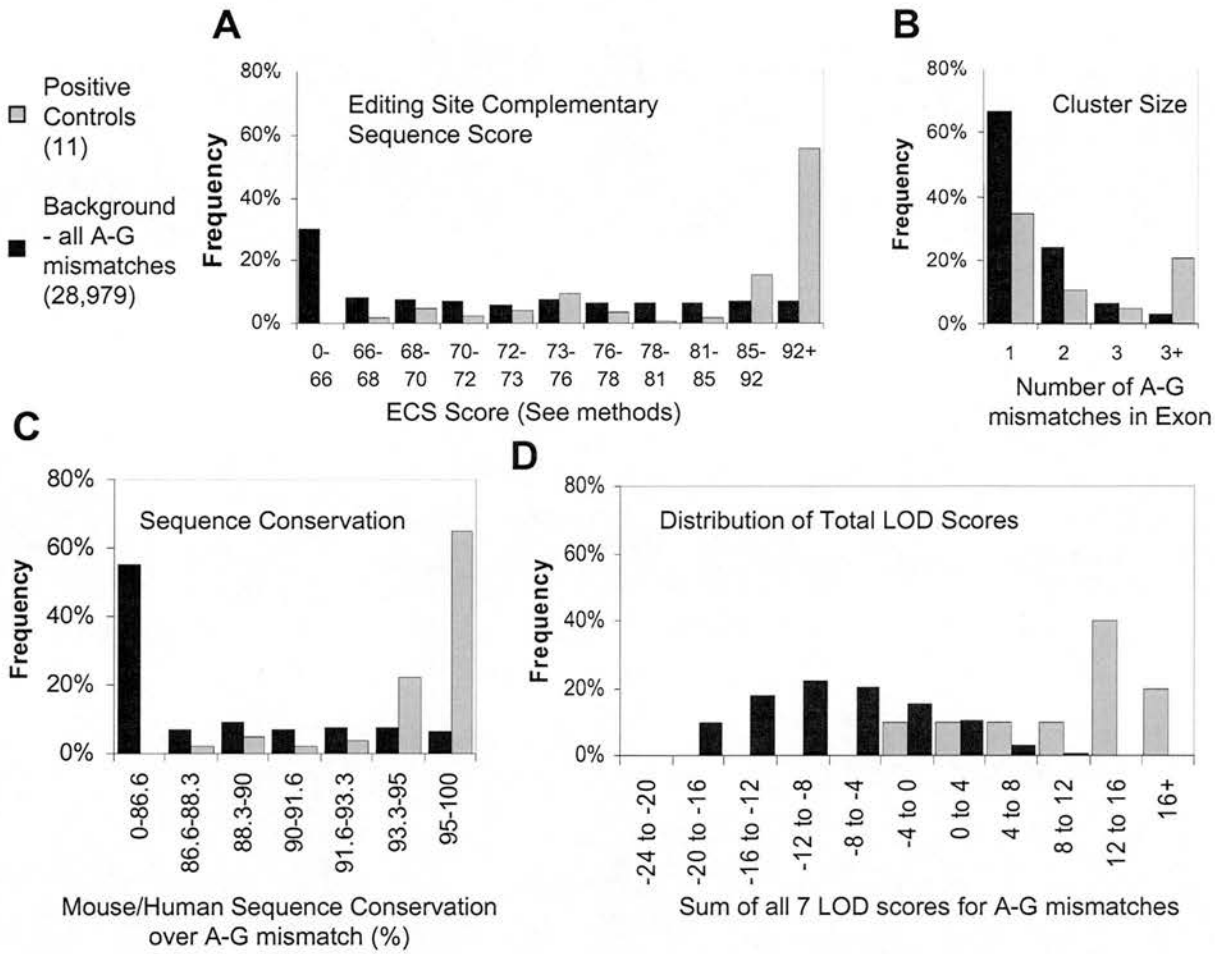
Each mismatch is annotated with its Ensembl gene ID, the overall LOD score, and its results for the seven tested features. The coding change of each of these A-G mismatches are also shown. The number of genomic sequences is the total number of genomic sequences aligned at the relevant position. Any positions where the genomic sequences vary were removed, so each of these will not contain the editing change. The final column identified the known edited sites and describes the results of the experimental validation (when performed).

Table 3.3. The Top 20 A-G Mismatches from the Mismatch-Based Screen (Section B)

Rank	Ensembl Mouse 30 Gene ID	A-Gs Observed In Brain	Repeat	Ensembl gene descriptions for Mouse and/or Human
1	ENSMUSG00000041380	1	No	5-HYDROXYTRYPTAMINE 2C RECEPTOR(SEROTONIN RECEPTOR)
2	ENSMUSG00000041380	1	No	5-HYDROXYTRYPTAMINE 2C RECEPTOR(SEROTONIN RECEPTOR)
3	ENSMUSG00000047214	8	No	BLADDER CANCER-ASSOCIATED PROTEIN (BC10).
4	ENSMUSG00000041380	2	No	5-HYDROXYTRYPTAMINE 2C RECEPTOR(SEROTONIN RECEPTOR)
5	ENSMUSG00000041380	2	No	5-HYDROXYTRYPTAMINE 2C RECEPTOR(SEROTONIN RECEPTOR)
6	ENSMUSG00000033981	5	No	GLUTAMATE RECEPTOR 2 PRECURSOR (GLUR-B) (AMPA 2)
7	ENSMUSG00000030776	0	No	TRINUCLEOTIDE REPEAT CONTAINING 6; EDIE
8	ENSMUSG00000001986	2	No	GLUTAMATE RECEPTOR 3 PRECURSOR (GLUR-C) (AMPA 3)
9	ENSMUSG00000025892	4	No	GLUTAMATE RECEPTOR 4 PRECURSOR (GLUR-D) (AMPA 4)
10	ENSMUSG00000020608	0	No	SMC6 PROTEIN
11	ENSMUSG00000033456	1	No	ADAPTOR-ASSOCIATED KINASE 1
12	ENSMUSG00000026014	0	No	AMYOTROPHIC LATERAL SCLEROSIS 2 CHROMOSOME REGION CANDIDATE 9
13	ENSMUSG00000021743	0	No	NA
14	ENSMUSG00000017776	0	No	PROTO-ONCOGENE C-CRK (P38) (ADAPTER MOLECULE CRK)
15	ENSMUSG00000030256	0	No	CLASS B BASIC HELIX-LOOP-HELIX PROTEIN 3 (BHLHB3) (HDEC2)
16	ENSMUSG00000026469	0	No	XENOTROPIC AND POLYTROPIC RETROVIRUS RECEPTOR
17	ENSMUSG00000049939	1	No	NAG14 PROTEIN.
18	ENSMUSG00000050587	1	No	NA
19	ENSMUSG00000024498	0	No	TATA BOX BINDING PROTEIN (TBP)-ASSOCIATED FACTOR
20	ENSMUSG00000004044	0	No	RNA POLYMERASE I AND TRANSCRIPT RELEASE FACTOR

Section B shows additional data for each of the top 20 A-G mismatches. This includes the number of A-G mismatches observed in brain or neural tissue and whether the edit occurs within a repeat sequence. Finally, the Ensembl gene description is given. If no mouse description was available, the human description was used.

Figure 3.2. Distributions of results for the three continuous variables and the final combined LOD score.



The seven positive control regions contain 10 recoding edits. The total number of all remaining A-G mismatches is 28,979. The four discrete variables are not shown here. In each of the following graphs, the distribution of the positive controls skews heavily towards the right. a) The ECS score describes the quality of the best predicted inverted repeat within 2kb of the mismatch (See Materials & Methods). b) The cluster size is the number of A-G mismatches in the exon containing the given mismatch. This graph does not include the erroneous GluR-D site (see Section 3.4.3). c) Sequence conservation over 120bp overlapping the mismatch. d) The sum of all seven LOD scores for each mismatch.

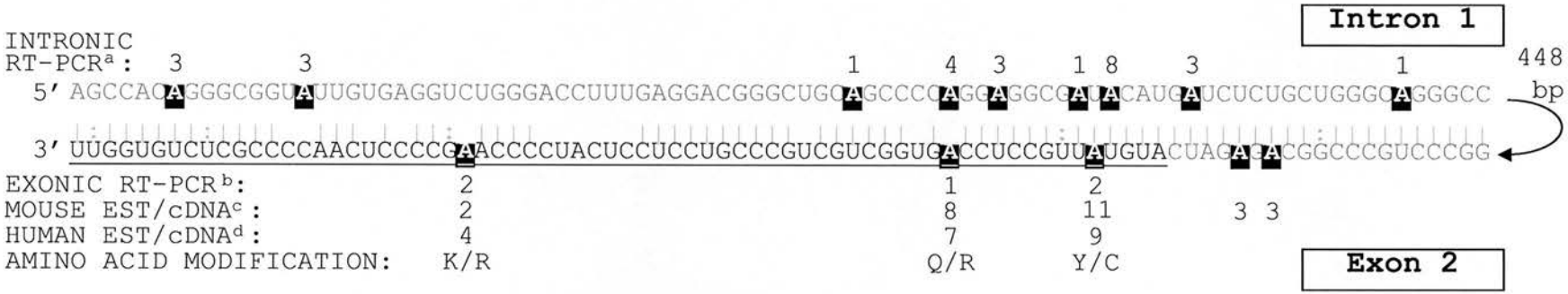
Using these data we can estimate the total number of mammalian A-I mRNA editing regions. Our protocol has identified six out of a total of ten known editing regions in the top ranked eight genes, representing 60% of the total. Given that *BC10* contains a verified editing region, this set includes seven genuine regions. These results imply that the total number of editing regions is roughly 12 (7/0.6). Given that ten regions are already known this suggests that there are very few mRNA recoding editing regions that remain unidentified. Based on this estimate, the proportion of all mouse exons that are edited would be 0.055% (12/21788). In addition, recoding editing appears rare compared to the total number of observed editing regions, accounting for less than 1% of the total number of editing sites identified[68,70]. Clearly these estimates are subject to error given the small sample size (see Table 3.1 - p46). However, assuming that the characteristics of known and unknown sites are similar, recoding editing sites appear to be rare in mammals. These data confirm previous suggestions that the majority of A-I editing occurs in UTR, intronic or repeat sequences[10].

In an effort to identify potentially interesting candidates we scanned the top 100 candidate edits to find any genes that were homologous to any of the known protein recoding edited genes in mammals or *Drosophila* (as predicted by Ensembl). With the exception of the known mammalian sites, no homologues were identified in this list.

3.4.5 Confirmation of *BC10* - A Novel A-I Editing Region

The top novel candidate, *BC10*, shows all the features of the positive controls. It has a very high scoring putative ECS 480bp 5' of the most frequently edited site, its edited region is 99% identical between mouse and human, it shows orthologous A-G mismatches in human, it is highly edited in brain tissue, affects the amino acid sequence and the editing sites occur in a tight cluster. The three recoding sites are supported by multiple ESTs/cDNAs in both species (see Figure 3.3 – p64). The putative ECS is the second highest scoring ECS from the ~30,000 A-G mismatches. The region containing these sites has been tested and editing has been verified in the lab. Interestingly, most of the editing was observed in the intron across the length of the predicted ECS and was specific to brain. The low number of edited RT-PCR products from the exonic region was partially due to an expressed pseudogene, which was preferentially amplified. *BC10* specific primers could not be found. This

Figure 3.3. Experimental Evidence for Editing in BC10



The predicted RNA duplex structure is shown with the putative A-I edited sites indicated by black boxes. The underlined nucleotides show the start of the coding region. The intervening 448bp are shown by the arrow on the left hand side. The supporting experimental evidence and the coding modifications of the putative edits are given above/below the putative sites.

a) The number of intronic brain RT-PCR products edited at this site (out of 50).

b) The number of exonic brain RT-PCR products edited at this site (out of 23).

c) The number of publicly available mouse brain expressed sequences edited at this site (out of 28).

d) The number of publicly available human expressed sequences edited at the orthologous sites (out of 85).

expressed pseudogene is not found in human and cannot explain the observed editing sites. These data demonstrate that this protocol is able to predict novel A-I editing sites.

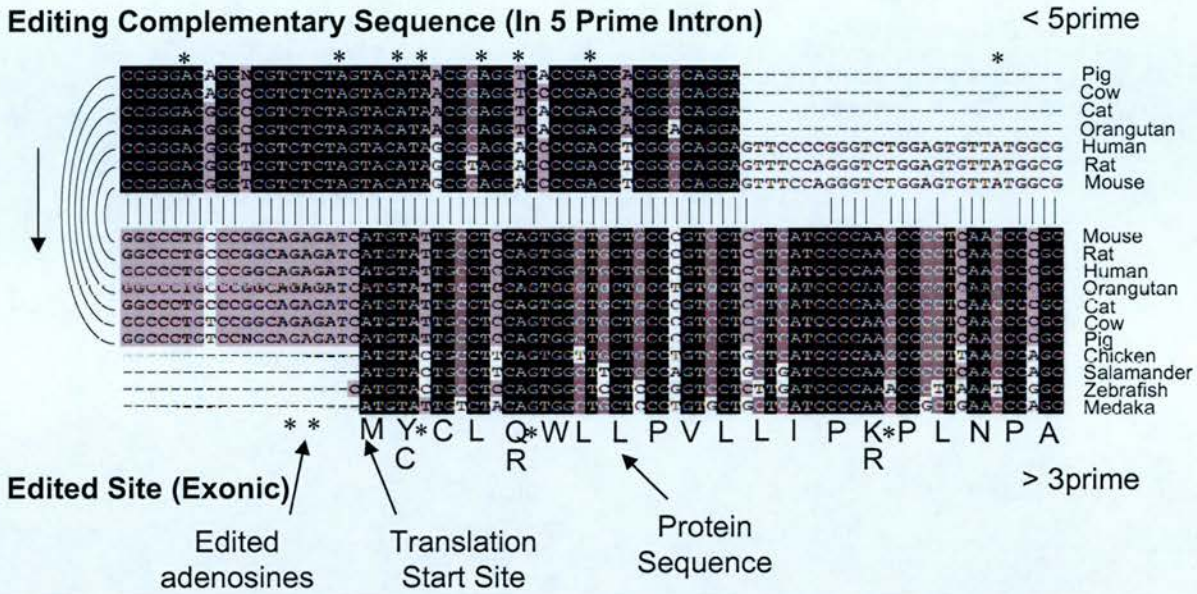
BC10 is predicted to be a small globular protein containing two transmembrane helices and is differentially expressed in bladder cancer[170] and renal cancer[171] cell lines. Gromova *et al* compared the mRNA expression patterns of invasive and non-invasive human bladder transitional cell carcinomas using differential display. Using this approach they identified BC10, which was exclusively expressed in non-invasive lesions. They suggested that down-regulation of BC10 may have a role in the transition from non-invasive to invasive bladder cancer. In contrast, Rae *et al* used differential display to compare renal-cell carcinomas and normal kidney gene expression. BC10 was one of 24 genes shown to be significantly up- or down-regulated. Further analysis confirmed that BC10 was down-regulated in renal-cell carcinomas.

All the editing sites are found in either the 5' UTR or the N-terminal section of the protein, which is predicted to be outside the membrane. The three coding edits are all non-synonymous and predicted to encode exposed residues. Figure 3.4 (p66) is a multi-species alignment of the *BC10* edited region that shows it is exceptionally well conserved from human down to fish, suggesting that it is a fundamentally important gene. Notably, the three recoding edited adenosines are conserved in all the species as well as most of the adjacent bases. This suggests that editing of this region may be conserved across all of these species.

3.4.6 An Editing Disease Gene?

Since this work was published, it has become apparent that amyotrophic lateral sclerosis (ALS) is linked to A-I RNA editing. It is not yet known what form this link takes, however, motor neuron cell death from increased Ca^{2+} influx due to reduced editing of the *GluR-B* receptor Q/R site provides one plausible hypothesis[39]. The authors suggest that this could be due to a general reduction in ADAR2 activity in motor neurons. Interestingly, the fifth best novel candidate reported here is in a gene called '*ALS 2 Chromosome Region Candidate 9*'. Although A-G mismatches were only observed in mouse for this site, it is well conserved (95.8% identity) and has a good inverted repeat structure. It is possible that this putative edited site is involved

Figure 3.4. Sequence Conservation of the BC10 Exon and ECS



DNA nucleotides entirely conserved across the ranges of species are highlighted in black. The ES is extremely well conserved between fish and mammals in the coding region. The 5'UTR is almost perfectly conserved between the mammals. The only mismatches in this region have compensatory changes in the predicted ECS. The predicted ECS is also highly conserved between mammals. There is no sign of the ECS in more distant species. Nucleotides edited in the RNA are indicated with a '*'. The protein sequence is shown next to the alignment. The amino acid changes induced by editing are also shown. The RNA base-pairing structure between the predicted ECD halves encoded by the DNA are shown between the aligned regions.

in causing this disease. Unfortunately, initial experimental validation failed to confirm editing at this site in adult brain and heart.

3.5 DISCUSSION

Our protocol is sensitive, given that it identified seven out of the ten positive control edits. In contrast no other screen for mammalian editing sites has identified any of the known recoding editing sites. Our protocol is also specific as five of the top eight genes contain known or experimentally verified sites. One of these is a confirmed novel editing region, *BC10*, which is an extremely well conserved gene implicated in renal and bladder cancer[170,171]. These results demonstrate that this is a useful and efficient method for identifying both known and novel A-I editing sites.

Using other computational protocols two groups have shown that there are over 1,500 genes edited in introns or non-coding regions[68,70]. This shows a clear difference in magnitude between the number of coding and non-coding editing sites. The results of Levanon *et al* [70] suggest that the occurrence of a strong ECS is a very good predictor of editing sites in non-coding sequences. Here we find that our ECS predictions are only moderate predictors for recoding sites and that a combination of features must be used to identify these sites. A better ECS prediction method could improve this situation. In contrast, we found that sequence conservation and the observation of orthologous mismatches in human are strong indicators of recoding sites. Notably, these features will not identify editing sites in Alu repeats, as they are not conserved in the mouse.

Despite the success of this protocol for identifying many of the known recoding sites, it is possible that there are many more sites to identify in addition to *BC10* and the previously known sites. For example, it is possible that some of the genes we experimentally tested may be edited, but that the degree of editing was too low, or restricted to particular tissues or developmental stages other than adult brain and heart. The frequency of editing is 100% for only the *GluR-B* Q/R site, whereas the frequency of editing for other sites is often much lower[1]. It can also be developmentally regulated as with the AMPA receptor R/G sites (*GluR-B,C,D*), which are poorly edited early in mouse development[32]. Although all the known mammalian recoding sites are edited in adult brain[1], there may be some

ascertainment bias towards brain editing. A large proportion of the publicly available expressed sequence data is from brain libraries, adding to this bias. Both the editing enzymes and edited Alu elements appear in a range of tissues[1,68,70], supporting the possibility of recoding editing in these tissues. Indeed the disease phenotype of *ADAR1* heterozygous mutations in humans is a skin condition, rather than neurological[38]. It is also possible that there is an additional class of recoding sites that do not conform to the features used in this analysis, however there is no evidence for this.

These data suggest that A-I recoding editing is rare in mammals, with an estimated total of approximately 12 editing regions. Given that we already knew of 10 editing regions in mammals, the *BC10* region appears to be one of only a few remaining edited regions. This result may have implications for any further screens for recoding editing sites.

Although our data suggest that recoding A-I editing is rare, the total number we predict is subject to several unavoidable sources of error, including the small size and relatedness of the positive control set. It is also possible that some of the genes we experimentally tested may be edited, but that the degree of editing was restricted to particular tissues or developmental stages other than adult brain and heart. Environmental factors, such as the disease state of the tissue, may also be important. One form of *ADAR1* is known to be interferon inducible[99], suggesting the existence of sites that are edited only during inflammation[100]. Aberrant A-I editing of the endothelin receptor has been implicated in Hirschsprung disease[82], while aberrant C-U editing has been shown to induce liver dysplasia and carcinomas[172,173]. It is possible that there are many more sites that could be aberrantly edited, some of which could contribute to disease. When looking for ECSs we identified more than 1,500 mismatches with inverted repeats that scored better than half of the known recoding sites. This suggests that there are a lot of potential hairpins formed between exons and their introns, which the editing enzymes could potentially bind. It is interesting to ask whether aberrant editing of these potential hairpins could be involved in disease.

Our results demonstrate that sequence conservation between mouse and human and the observation of orthologous mismatches are powerful predictors of recoding editing sites. As a result, this scoring system will be less useful for identifying

putative species-specific sites (i.e. mouse specific sites). However, most of the positive controls have been shown to be widely conserved throughout mammals[1].

In addition to recoding sites, there are many non-coding sites remaining to be discovered or characterised[68,70]. Understanding the functions of these non-coding sites is vital for a fuller understanding of RNA editing. This work has identified many of the known mammalian recoding editing sites and one novel edited region in *BC10* and it is clear that there may be further sites to be identified. The ALS candidate region gene is of particular interest. However, the present data suggest that recoding editing is a rare phenomenon, both as a proportion of total editing activity and as a proportion of affected exons in the mammalian genome.

4 Results: Conserved RNA Duplexes in Vertebrates

4.1 Preface

The previous chapter demonstrates that a screen based on mismatches can successfully identify novel RNA editing sites. However, it also highlights the potential problems with this approach. Firstly, the degree of noise from SNPs, sequencing errors and mis-alignments make the majority of predictions unreliable. Secondly, there are many known editing sites that could not be distinguished in this data set. The main reason for this is that there were very few or no edited sequences in the public databases. This limitation of the expressed sequence databases implies that there may be many more sites that have not been identified due to poor coverage. This will be especially true of genes that are edited at low frequency, in poorly sampled tissues, in poorly sampled life stages, or are expressed at a low level. These issues become more important for species other than mouse and human, for which the number of publicly available expressed sequences is substantially smaller.

For these reasons it was decided to attempt a screen that does not rely on mismatch data. Instead, this screen was based on the identification of putative RNA duplexes that are conserved in more than one species. This chapter describes an analysis of and screen for editing complementary duplexes (ECDs) in a range of vertebrates, including mouse, rat, human, chicken, pufferfish and zebrafish.

4.2 Introduction

The requirement of a double stranded RNA (dsRNA) structure has been demonstrated for many of the mammalian edited sites[2,23,32,78,80,85]. Figure 4.1 (p75) shows the known dsRNA structures (termed ECDs) that have been published for the known mammalian protein recoding edited sites. Disruption of these dsRNA structures has been shown to reduce or completely stop editing[10]. An imperfect duplex has been shown to be a requirement for specific editing[77]. The duplexes for these edited sites vary in size, but each has a strong core of roughly 20-30bp, that is both well conserved and well base-paired (personal observations). These core ECDs

are shown in Figure 1.4 (p12). The numbers of unaligned bases, gaps or bulges in these dsRNA structures also varies.

The established methods for identifying the ECSs for known edited sites are not very satisfactory. Some of the known ECSs were identified through a time-consuming method of scanning by eye. This method is particularly difficult as it is difficult to take into account RNA base pairing specificities or the inclusion of gaps or bulges. For some of the edited genes, MFOLD[174] has been used to predict possible RNA structures, which would indicate the position of the ECS. Unfortunately, these programs can be unreliable, especially when only given part of a pre-mRNA to fold[175]. This is typically the case as the computing power required to fold an entire pre-mRNA is extremely prohibitive. This method does appear to work for some of the ECSs however[81]. Another method that also has had some success is the application of phylogenetic analysis to identify regions that are very highly conserved across a range of species. The three glutamate receptor R/G sites show an exceptional degree of conservation between mammals and fish[85]. However, there are still many editing sites for which ECSs have not been identified or published.

In the previous chapter a relatively simple method for identifying putative ECSs was presented. This method used a local alignment algorithm with RNA base-pairing specificities to look for potential double stranded structures. This analysis was very helpful for identifying the genuine editing mismatches, however, it did not have sufficient resolving power to be used on its own (see Table 3.2 – p57). The main reason for this was that the alignment scores obtained for the positive control ECSs were not sufficiently high enough to be rare in the genome; i.e. many supposedly non-edited exons had putative ECSs scoring as high or greater than many of the positives.

This first part of this chapter describes how this simple method was improved, primarily by combining it with a comparative genomics approach. The second part of this chapter describes how the improved method was applied to the known edited genes. Finally, the third part of the chapter describes how this method was used to screen a series of vertebrate genomes for novel edited genes.

4.3 Improving the ECS Search Specificity

A number of changes were made to the original ECS search method. Previously, the location of the putative edited site was taken to be the sequence surrounding the A-G mismatch being analysed and it was only the ECS that was unknown. The search method used in this chapter identified both halves of the ECD. The main improvement to this ECD search method is the addition of a comparative analysis of the ECDs between two species. One requirement of this change was to use a different local alignment algorithm (see Materials & Methods).

4.3.1 A New Local Alignment Algorithm

Previously, the Water algorithm from EMBOSS was used to identify putative ECSs. This has the disadvantage that it only predicts a single match between the query and target sequences (the target sequence for this method is simply the reverse of the exon and surrounding sequences). This was not a problem for the previous protocol as only the score of the best match was of interest. The protocol in this chapter, however, was interested in the best *pair* of overlapping ECD predictions between two species. This pair will not necessarily consist of the top scoring ECD predictions from both species. The results for some of the known edited sites demonstrate that this is indeed the case. Matcher/LALIGN is a local alignment program that predicts multiple local alignments, and reports more than one match. The matrix used by Matcher/LALIGN was the same as the one used by Water in Chapter 3. This matrix was designed to allow for the ability of uridine to base pair with both adenosine and guanosine in RNA secondary structures.

4.3.2 Comparative Analyses of the Putative ECSs

A striking feature of the known edited sites and their ECS sequences is that they both tend to be very well conserved between species. In contrast, most of the other putative ECDs that are predicted for the known edited exons tend not to be so well conserved (personal observations). The method described here uses this observation to identify putative ECDs with considerably higher specificity than an analysis based only on one species.

So instead of looking for a single high scoring ECS, this method looked for pairs of overlapping ECSs between two species. The assumption was that if two putative ECSs have been conserved then they are more likely to be part of a functional ECD. However, it should be noted that this does not necessarily mean that they are edited, as there are other explanations for conserved duplexes. The exact methods and requirements used for identifying ECDs are given in the full protocol description.

4.3.3 Where To Look

The majority of the published mammalian ECS sequences are found in the adjacent 3 prime intron (see Figure 1.4 – p12). There are also published ECS sequences found in the adjacent 5 prime intron (e.g. *BC10* & *ADAR2*) and in the exon itself (e.g. *KCNA1*)[63,79-81]. The distance between the two halves of the known ECDs in the mouse ranges from zero for the *GluR-B* R/G site[32] to 1.8kb for the *GluR-6* Q/R site[78]. Table 4.1 (p73) shows that the majority of ECSs occur within 500bp of the ES. This bias may have occurred because more distant ECSs are harder to identify.

Table 4.1. Separation of the ECD Halves in the Known ECD Structures.

Distance Between ECD Halves	Number of Known Mouse ECDs	Sites
0 - 0.5 kb	6	*GlurR-BC&D* R/G, *GluR-B* Q/R, *5-HT2C*, *KCNA1*
0.5 – 1.0 kb	0	-
1.0 – 1.5 kb	1	*ADAR2*
1.5 – 2.0 kb	2	*GluR-5* Q/R, *GluR-6* Q/R
>2.0 kb	0	-

Given these observations, separate searches for ECDs were performed in the exon itself and in the flanking 2.5kb of each adjacent intron. If another exon occurred within 2.5kb in either direction of the exon being analysed, then the whole intron sequence was used instead. These criteria incorporated all the known ECSs, but would not have incorporated any novel sites that occur further away. A disadvantage of searching further away was that the probability of ECS structures being observed by chance increased in proportion to the length of sequence searched.

All of the resultant ECDs had one half in the exon. The locations of the other halves (the ECSs) were in the 5 prime intron, the 3 prime intron, or in the exon. High sequence conservation in introns is considerably more rare than in exons, especially if they are coding exons. By extension, conserved inverted repeats would also be more common. This means that ECDs where both halves occur in the exon are generally less significant than if the same structure had been observed between the exon and an intron. This is the main reason for treating ECS searches in introns and exons separately.

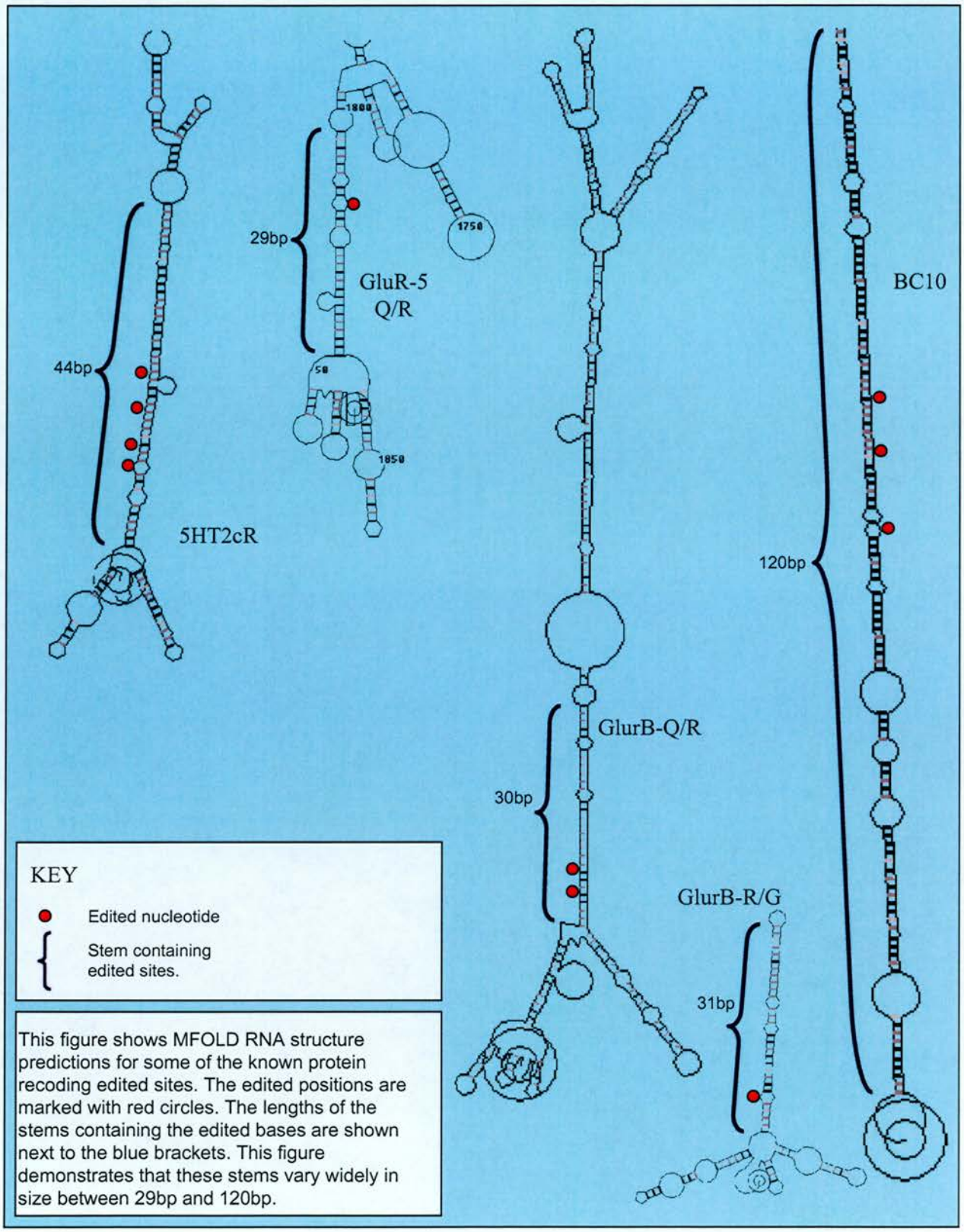
For several of the known edited sites, the published ECD structure shows the majority of the ES half of the ECD to be outside the exon, as shown in Figure 1.4 (p12). For this reason an extended version of the exon sequence, with an additional 50bp on either side, was used in the searches to identify ECDs. This allowed the method to identify ECSs that occur at or overlapping the exon/intron boundaries.

4.3.4 Looking for a Minimal ECS

During the development of the ECS search program, a large number of measurements were taken to gauge the program's success. These included exonic sequence conservation, intronic sequence conservation, secondary structure quality scores for both species, and conservation of secondary structure between species. These scores were highly dependent on the length of the predicted inverted repeats in each species, as well as the length of the overlap between them. Unfortunately, these lengths varied significantly between the predictions for the editing sites. For this reason it was difficult to meaningfully compare results between different editing sites.

MFOLD RNA structure predictions were generated for the known edited sites. These showed that the edited sites all occur in RNA stems with a minimum size of just under 30bp, although some are much longer. Figure 4.1 (p75) shows some of these MFOLD predictions. Based on this observation, it was decided to modify the search to identify conserved 25bp ECDs. This is a liberal choice of ECD length, in that it should have allowed the identification of all the known ECDs, as well as being able to predict any that are slightly shorter. It is important to note that this length is not necessarily a requirement for ECDs, so much as a reasonable, minimal

Figure 4.1. MFOLD RNA Structure Predictions for Known Edited Sites



This figure shows MFOLD RNA structure predictions for some of the known protein recoding edited sites. The edited positions are marked with red circles. The lengths of the stems containing the edited bases are shown next to the blue brackets. This figure demonstrates that these stems vary widely in size between 29bp and 120bp.

approximation that was based on the known examples. This removed the complications induced by predicted inverted repeats with different lengths and different overlaps, and created a standardised method to compare results. Henceforth, I refer to the best 25bp window as the putative core ECD.

A scoring system for the putative ECDs was generated, which measured the degree of sequence change in the ES and ECS sequences as well as the quality and conservation of the secondary structure. This system is detailed in the full protocol description (Section 4.4).

4.3.5 Using Multiple Species

So far, the method has only been described for a comparison between two species. However, many of the published ECDs can be seen in a range of metazoa. For example, the *GluR-C* R/G ECD can be seen in mouse, rat, human, pufferfish and zebrafish[85]. In this example it would be considerably more effective to look for an ECD that is conserved between all these species, instead of just mouse and human. In practice it is not trivial to carry out multiple alignments. This becomes even more complicated when you try and align alignments, such as those provided by the ECD predictions, instead of simple sequences. However, many so-called 'multiple alignment' programs cheat by doing a series of pair-wise alignments. This approach was applied to this protocol.

Using mouse as a base species, pair wise comparisons were made with the orthologous exons from rat, human, chicken, pufferfish and zebrafish. These model organisms represent a range of evolutionary distances up to 450million years from mouse[86]. ECDs that were observed in more than two of the species were considered more convincing than those that were only seen between two species. In practice, this resulted in a significant increase in the specificity of the ECS search (see Section 4.5.1). Mouse was chosen as the base species as it was relatively easy to perform laboratory verification of any predicted novel ECDs in this model organism.

4.4 Full Protocol Description

4.4.1 Data Preparation

4.4.1.1 Initial Files

A number of files from external sources were required for these analyses. These are detailed in this section. The following files were obtained for mouse, rat, human, chicken, pufferfish and zebrafish;

- Genomic nucleotide sequence files from Ensembl[142] (see Materials & Methods). The versions and types of files are:
    - Mouse : NCBI Build 33 (Obtained November 2004). 21 Full chromosomal sequences (including 1-19, X & Y).
    - Rat : Baylor RGSC3.1 Build (Obtained November 2004). 21 Full chromosomal sequences (including 1-20 & X). Additional file of concatenated unmapped segments.
    - Human : NCBI Build 35 (Obtained November 2004). 24 Full chromosomal sequences (including 1-22, X & Y). Additional file of segments not mapped to a chromosome.
    - Chicken : WASHUC1 Build (Obtained November 2004). 30 Full chromosomal sequences (including 1-24, 26-8, 32, W and Z). Additional file of sequences not mapped to a chromosome.
    - Pufferfish : FUGU2 Build (Obtained November 2004). A 5.4X whole genome shotgun assembly of 4.1 million fragments, assembled into 20379 sequence scaffolds.
    - Zebrafish : Sanger WTSI Zv4 Build (Obtained November 2004). 25 chromosomal sequences (including 1-25). Additional file of sequences not mapped to a chromosome.
- All predicted exon sequences & coordinates from Ensembl version 26 (see Materials & Methods).
- All gene descriptions from Ensembl version 26 (see Materials & Methods)

The following files were obtained for mouse vs. rat, human, chicken, pufferfish and zebrafish;

- Orthologous gene predictions from Ensembl version 26 (see Materials & Methods).

The genomic sequence files were formatted into BLAST databases using the formatdb program that comes with the BLAST package (version 2.2.6). Additional sequence indexes were generated for both the genomic and the exon FASTA files (as required by Sgrab – see Section 2.51). This allows for rapid sequence retrieval by the programs that use this data. The method behind this rapid retrieval is described in the Materials & Methods.

4.4.1.2 Orthologous Exon Predictions

Although orthologous gene lists have already been obtained from Ensembl, it was not clear which exons were orthologous to each other. This information was obtained, however, by BLAST searching the exons against each other. As the whole genome was not being searched, reciprocal BLAST analysis was not required. Any good match was considered to be a putatively orthologous exon.

Each mouse exon was BLAST searched against a BLAST database of all the exons in the orthologous gene. The BLAST options used are (-F F –m 8 –e 10 –r 1 –q –1 –G 2 –E 1 –W 9). These options removed the simple sequence filter, modified the match, mismatch and gap penalties, shortened the word length to 9 and restricted the output to tabular results with E-values less than 10. These settings allowed for a more sensitive, albeit slower, BLAST search between the two sequences. Several filters were then applied in an effort to ensure that only genuine orthologous exon predictions were generated. Exons matches shorter than 10bp or with a BLAST bit-score less than 50 were ignored. The BLAST match nucleotide identity percentage had to be above suitable thresholds for a given species. These were as follows; rat – 85%, human - 80%, chicken – 70%, pufferfish – 60% and zebrafish – 60%. These values were chosen as they find most of the clearly orthologous exons, while reducing the number of predictions that appear to be erroneous.

One outcome of this analysis was that a single mouse exon could have multiple predicted orthologous exons in each species. This non-conservative approach allowed a more thorough analysis of all possible orthology relationships, although this also introduces some unwanted noise.

4.4.2 Main Program

The main part of this protocol was incorporated into a single program, entitled 'scan_orths_multi.pl'. The input for this program is a list of putative orthologous exons and a number of variables (which are discussed at the end of this section). The following instructions were carried out for every putatively orthologous pair of exons. Figure 4.2 (p80) is a flowchart that outlines the major steps involved in this protocol.

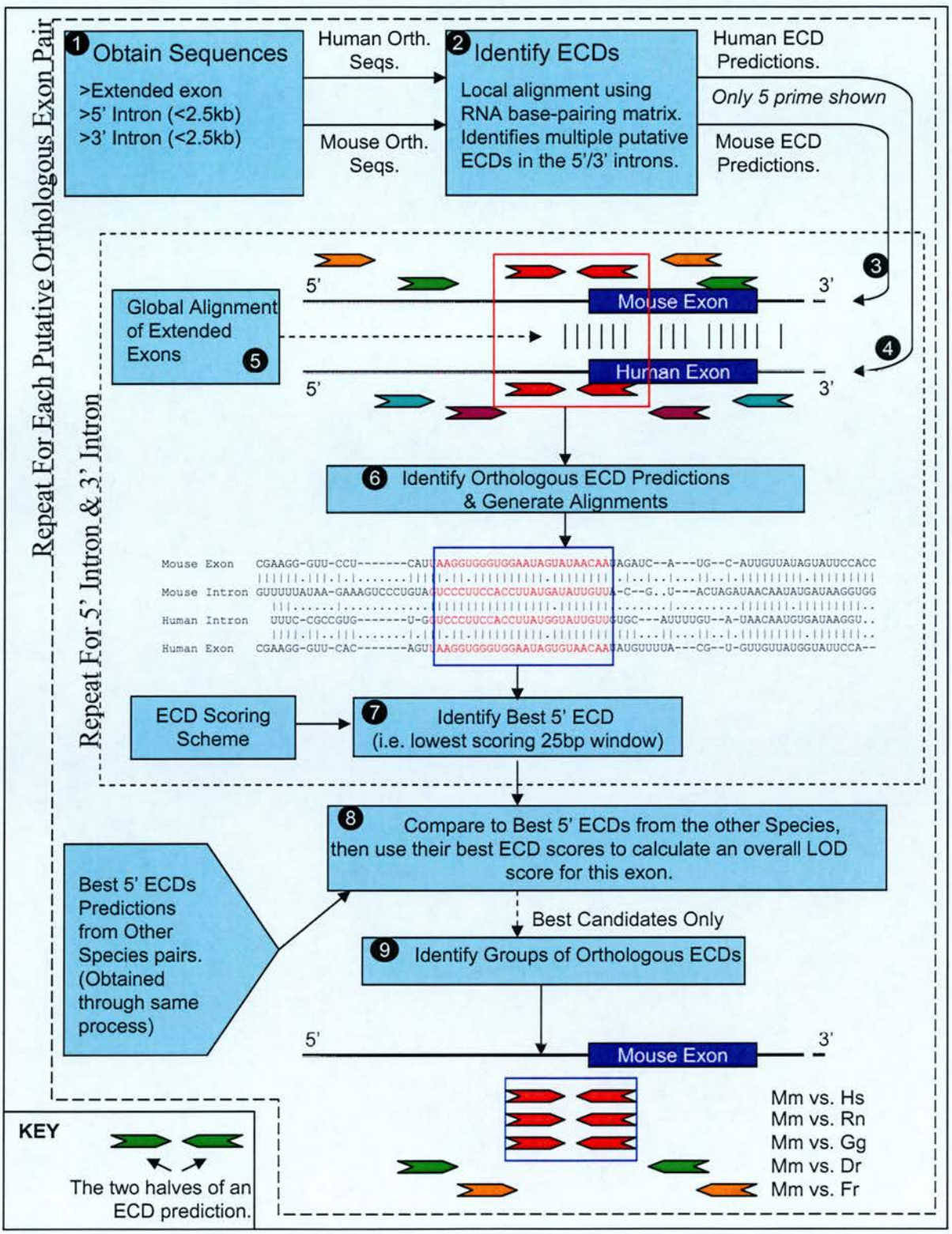
4.4.2.1 Preparing the Sequences

The exon, 5 prime intron and 3 prime intron sequences were obtained in FASTA format for both species being analysed. The intron sequences had a maximum length defined in the variables (default was 2.5kb), but were shorter if there was an exon on either strand before this length was met. This restriction ensured that the extremely long introns did not introduce an unwanted source of noise. Extended exons, with an additional 50bp flanking either side, were also obtained. These sequences were all obtained using the rapid sequence retrieval system described in the Materials & Methods. Where the sequences are found on the reverse strand, they were reverse complemented.

4.4.2.2 Searching for ECSs

Using a local alignment algorithm and the RNA base-pairing specificity matrix described and derived in Chapter 3, it was possible to identify simple RNA duplex structures. In this protocol, the alignment algorithm used is LALIGN (see Materials & Methods). The alignments were carried out between the extended exon sequences and their respective 5 prime intron, 3 prime intron or exon sequences (i.e. extended exon vs. the same exon sequence). In each case the second sequence was reversed to reflect the conformation of RNA duplexes. The options used with LALIGN were '-f –12 –g –3 –s rna_matrix –m 10 file1 file2 20', where file1 and file2 were the two sequences being aligned. These options specified a gap-opening penalty of –12 and a gap-extension penalty of –3, and specified that the algorithm should result in 20 alignments.

Figure 4.2. Flowchart for Identifying Conserved ECDs



Mm = Mus musculus (Mouse)	Rn = Rattus norvegicus (Rat)	Hs = Homo sapiens (Human)
Gg = Gallus gallus (Chicken)	Dr = Danio rerio (Zebrafish)	Fr = Fugu rubripes (Pufferfish)

Figure 4.2. Flowchart for Identifying Conserved ECDs (Continued)

LEGEND: This flowchart provides an overview of the protocol described in this chapter. The process has been divided into nine steps, each of which is repeated for every orthologous exon pair, in each pair of species (Mm vs. Rn, Mm vs. Hs, Mm vs. Gg, Mm vs. Dr, Mm vs. Fr).

Step 1. *Sequences are obtained.* These include the exon, and extended exon (with 50bp flanking sequence), the 5 prime intron (up to 2.5kb) and the 3 prime intron (up to 2.5kb). These sequences are obtained for both species (e.g. Mouse and human). Each step after this is performed for both introns.

Step 2. *Identify ECDs for Each Species.* A local alignment algorithm, with an RNA base-pairing matrix, is used to identify putative RNA duplexes between the extended exons and their intronic sequences. Many ECDs are predicted per species.

Step 3. *Align the mouse ECD structures to the mouse sequence.*

Step 4. *Align the ECD structures from the second species to the second sequence.*

Step 5. *Align the two extended exons so that the locations of the ECDs can be compared.* This is performed using a standard DNA global alignment program.

Step 6. *Identify Putative Orthologous ECDs and Generate Alignments.* An example of this is shown in the red box. At this stage ECDs were considered to be putatively orthologous if the ES halves overlapped each other by 25bp or more. An alignment is then generated by combining the two local alignments that resulted in the individual ECD predictions.

Step 7. *Scan the Alignment for the Best 25bp Window.* Each window of 25bp in the alignment is scored for both the sequence conservation and the quality of the secondary structure. The lowest scoring window is the best.

Step 8. *Compare the Best ECDs from Each Species for this Exon.* The scores of the best ECD prediction from each species and location (5 prime or 3 prime) are converted into LOD scores. In contrast to the original scores, the LOD scores can be summed to fairly generate an single overall score for the exon based on the ECDs in all the species pairs.

Step 9. *Identify Groups of Orthologous ECDs.* The alignments are manually checked to confirm that the predicted ECDs overlap each-other. This results in the identification of groups of orthologous ECD predictions.

It was also possible to search for ECDs where the ECS is in the exon itself. However, the randomisation approach used does apply to these ECD predictions so they have not been shown on this diagram.

4.4.2.3 Finding Orthologous ECS Predictions

The putative ECDs predicted by this algorithm for each of the two species needed to be aligned to see if they overlap (i.e. both halves occupy orthologous positions). For this purpose a global alignment of the exons was carried out using the LAGAN algorithm (default options). This alignment was then annotated with each of the predicted ECDs from both species. This alignment was scanned to identify any mouse ECDs that overlap a predicted ECD from the second species by specified amounts (default is 25bp for both halves). This resulted in a list of putatively orthologous pairs of ECD predictions.

4.4.2.4 Scoring the Core ECS Predictions

A sliding window, with each window containing exactly 25bp of mouse exon, was used to find the best 'core' segment of the ECS pair alignment. The best core segment was the one with the lowest score. This score was based on exon conservation, intron conservation and secondary structure quality and conservation. This scoring scheme was designed so as not to discriminate too strongly against gaps or mismatches caused by bulges or small loops in the putative hairpins. The details of the scoring scheme were as follows;

- Exon conservation
    - Score is +0 if the two ES sequences are the same at a given position.
    - Score is +1 for each of the first two mismatched or gapped nucleotides.
    - Score is +0.5 for additional mismatches or gaps.
- Intron conservation
    - Score is +0 if the two ECS sequences are the same at a given position.
    - Score is +1 for each of the first two mismatched or gapped nucleotides.
    - Score is +0.5 for additional mismatches or gaps.
- Secondary structure
    - Score is +0 if both ECDs base pair at this position.
    - Score is +1 for each of the first two positions when neither ECD is base-paired.
    - Score is +0.5 for additional positions when neither ECD is base-paired.

- Score is +2 for each of the first two positions where one ECD is base-paired and the other is not.
- Score is +0.5 for additional positions where one ECD is base-paired and the other is not.

Compensatory changes incurred a conservation score penalty, but as base pairing was maintained, they did not incur a score penalty for secondary structure.

This scoring scheme was trained and tested on the published ECSs of known editing sites. It is not a trivial task to determine the optimal scoring scheme for this task; however, the method provided appeared to work sufficiently well. Once each score was calculated, the putative core ECD alignments were recorded for future observation and analysis. Figure 4.2 (p80) provides an example of an ECS alignment with the core section highlighted in red (between steps 6 & 7).

4.4.2.5 Program Options

There are a number of options associated with this program, each causing fundamental differences to the operation of the program. These are listed below;

Orthologous Exon File

> This program can be pointed at any file containing orthologous exon predictions (in the correct format), allowing the analysis of specified sub-sets of genes or exons.

Restricted Analysis

> The program can be told to ignore all orthologous exon predictions unless they contain a defined text string. For example, this could be an Ensembl gene or exon identifier.

Intron Scan Distance

> This variable allows the user to define the length of introns to be searched. By default this is 2.5kb.

Core ECS Size

> This variable allows the user to define the length of the core ECS window. This allows the user to look for ECS predictions of different lengths.

Make Control Introns

> This variable is required to generate control introns, which are required for the statistical analysis. Details of this are covered in the following sections.

Run with Control Introns

This variable is used to specify that the program should use the control introns derived using the previous command, instead of the real introns associated with each exon.

Output Format

The output format can be modified to include all the ECS predictions or just the best, as well as deciding between tabular results, full alignments or both.

Help

This variable results in the program presenting brief help message to inform the user of the available options and their defaults.

4.4.3 Analysis of the Putative Conserved ECSs

Using the above scoring system to analyse all orthologous exons resulted in a large amount of data. This data contained both genuine ECD predictions and presumably considerable amounts of false predictions.

It was important to gauge the significance of obtaining any given putative ECD and there are two methods that were considered. Firstly, by comparing a given ECD to a distribution of the known ECD scores, versus all remaining predicted ECDs, we could have calculated the probability that it belonged to either distribution. The assumption here is that the 'remaining' distribution may have contained some genuine unknown ECDs, but these will be so rare that they should not affect the results. This method would have directly assessed how likely an ECD was to be genuine. Unfortunately, the number of known conserved ECD structures in vertebrates was too low to perform this with any accuracy.

The second method was to use a randomisation approach, which removed the need for a distribution of known ECDs. Instead each orthologous exon pair was paired with a set of four introns from another randomly selected orthologous exon pair (one from both introns in both species), and a randomised distribution of top-scoring ECD predictions was generated. By comparing this to the real distribution, it was possible to say how likely a given ECD would be by chance. This method did not directly assess how likely an ECD was to be genuine. Instead it indirectly assessed how likely the putative ECD structure was to be functional. Each ECD structure could have been conserved for reasons other than editing, such as splicing[122]. Although this is

indirect evidence, it was the method that was chosen. The randomised distribution was generated using the 'make control' variable in the main program (see above).

This scoring scheme allowed us to provide each top-scoring ECD with a meaningful score. However, we wanted to combine the top-scoring ECD scores between species in order to give the analysis more power. For this purpose we used a relative entropy scoring system (see Materials & Methods). This allowed us to combine the ECD predictions from separate species comparisons, as long as they were independent. Due to the similar evolutionary distance between mouse and the two fish (zebrafish and pufferfish), we combined their results, such that only the best scoring ECD from either pair of species was considered. This was applied to both the real and the randomised distributions and LOD scores were generated for these combined distributions.

4.4.3.1 Relative Entropy and LOD Scores

The relative entropy was calculated as a log-of-odds score, based on the following formula[162]. For a given pair of species and an ECD with score x, the LOD score (L) is;

$$L_x = \log_2\left(\frac{Rl(x)}{Rd(x)}\right)$$

where Rl(x) is the proportion of the real distribution in an interval containing score x, and Rd(x) is the proportion of the randomised distribution in the same interval. On the assumption that the ECD scores between each pair species are independent, they can be summed. This allows us to compare each putative ECD based on its conservation across a range of species.

In reality the observation of an ECD in one pair of species will not be independent from observations of the same ECD in another pair of species. This is because both comparisons used mouse as a base sequence. It is not clear how this bias could have been avoided. However, the results of this scoring scheme demonstrated that while it may not be fully statistically rigorous, it was effective.

4.4.3.2 Practicalities of the Relative Entropy System

Figures 4.3–4.6 (p87) show how the two real and randomised distributions compare for each pair of species. Only ECDs predicted to form between the exon and either intron are shown in these graphs. As the randomisation procedure relied on swapping the intronic sequences, it is not appropriate for the prediction of ECDs that have both halves in the exon.

The best ECDs are those with the lowest scores. Section A of each figure shows the overall cumulative distributions of the real and random ECD scores. In the most distant species pairs the two distributions are largely very similar, although the real distribution generally scores slightly better. These observations suggest that there is not much more conserved secondary structure than observed between these distantly related species. The closer the two species are, however, the greater the difference between the distributions. For the mouse:rat distribution, there is a large difference between the real and random distributions. This suggests that there may be a relatively large degree of structure conservation between mouse and rat. Both the real and random mouse:rat distributions derive from a high proportion of high quality/low scoring ECDs. As these two species are so closely related, their sequences are generally highly conserved, which means that almost *any* duplexes that form are likely to have low scores.

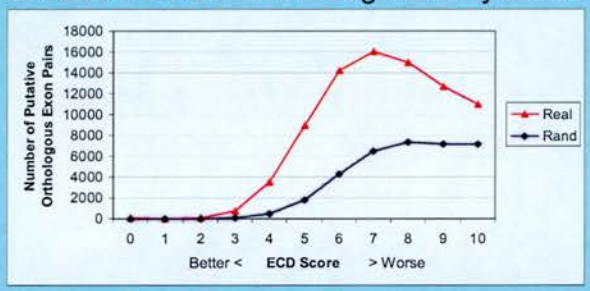
Section B of each figure focuses on the high quality/low-scoring ECDs from each distribution. In each figure the real distribution consistently lies above that of the random distribution, which suggests that many ECDs at this end of the real distribution have some reason for being conserved – i.e. they are functional. The number of ECDs with low scores in each species pair decreases rapidly with divergence between the species pair. One effect of this is that the distributions become considerably more erratic at lower scores for the more distant species (i.e. chicken and fish).

Section C shows LOD scores derived from these distributions. These have been calculated for each score where at least 3 putative ECSs have been found in both the real and randomised score bins. This reduces the error in the calculated LOD scores. For the mouse:rat, mouse:human and mouse:chicken comparisons, these LOD scores show relatively consistent trends. The LOD scores for fish, however, are more erratic due to the low numbers of both real and random ECD predictions of high quality. In

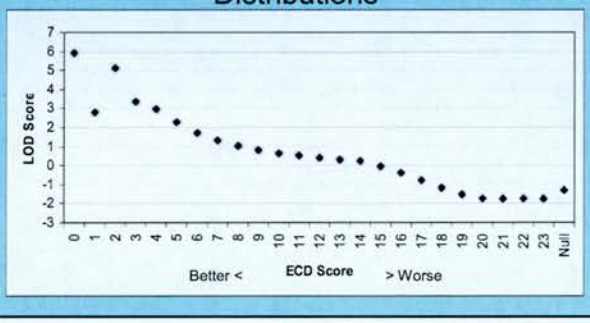
Figure 4.3. ECD Score Distributions and LOD Scores for Rat

A. Cumulative Proportion of ECDs Below Score Threshold

B. Absolute Numbers of High Quality ECDs



C. LOD Scores Derived from these Distributions



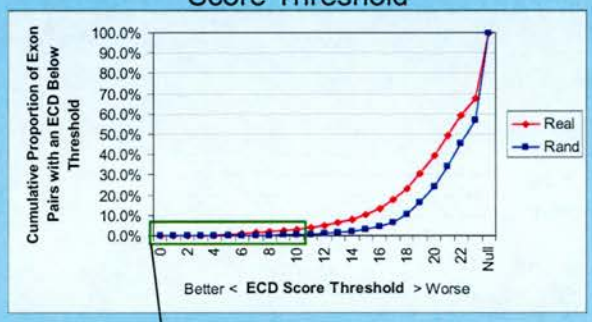
Section A shows the cumulative distributions of all mouse:rat real and randomly generated ECD predictions. The green box indicates the region that is expanded in Section B.

Section B shows the absolute numbers of real and randomly generated ECD predictions with ECD scores less than or equal to ten. These are the highest quality ECD predictions.

Section C shows the LOD scores generated from a comparison of these distributions. LOD scores are only given where 3 or more ECDs have been predicted in both the real and random distributions.

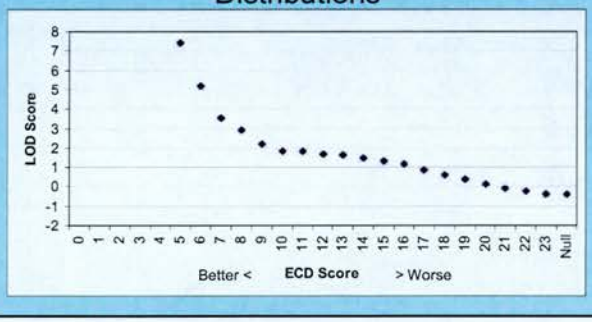
Figure 4.4. ECD Score Distributions and LOD Scores for Human

A. Cumulative Proportion of ECDs Below Score Threshold



B. Absolute Numbers of High Quality ECDs

C. LOD Scores Derived from these Distributions



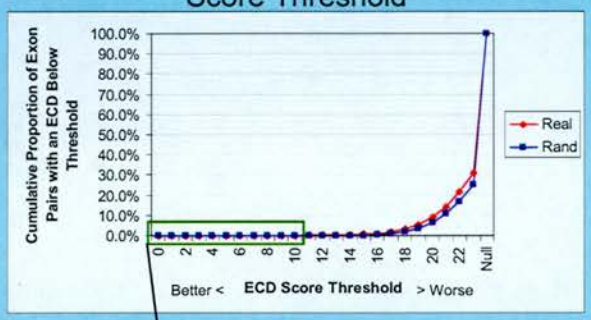
Section A shows the cumulative distributions of all mouse:human real and randomly generated ECD predictions. The green box indicates the region that is expanded in Section B.

Section B shows the absolute numbers of real and randomly generated ECD predictions with ECD scores less than or equal to ten. These are the highest quality ECD predictions.

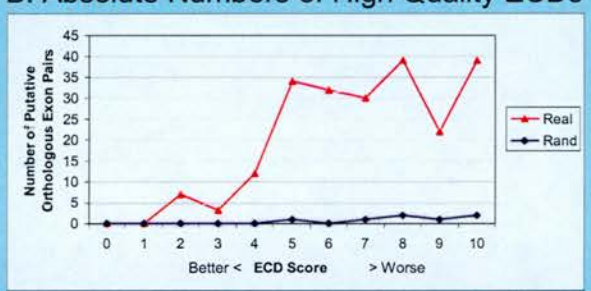
Section C shows the LOD scores generated from a comparison of these distributions. LOD scores are only given where 3 or more ECDs have been predicted in both the real and random distributions.

Figure 4.5. ECD Score Distributions and LOD Scores for Chicken

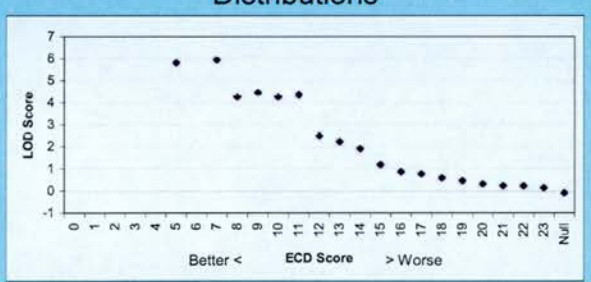
A. Cumulative Proportion of ECDs Below Score Threshold



B. Absolute Numbers of High Quality ECDs



C. LOD Scores Derived from these Distributions



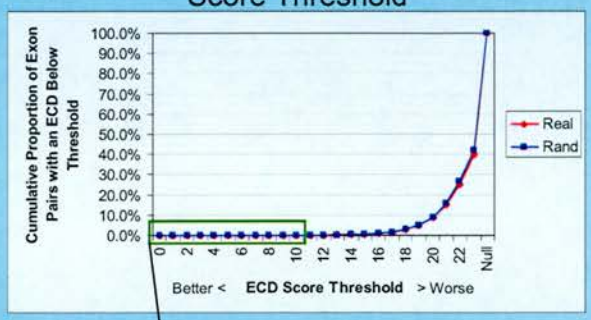
Section A shows the cumulative distributions of all mouse:chicken real and randomly generated ECD predictions. The green box indicates the region that is expanded in Section B.

Section B shows the absolute numbers of real and randomly generated ECD predictions with ECD scores less than or equal to ten. These are the highest quality ECD predictions.

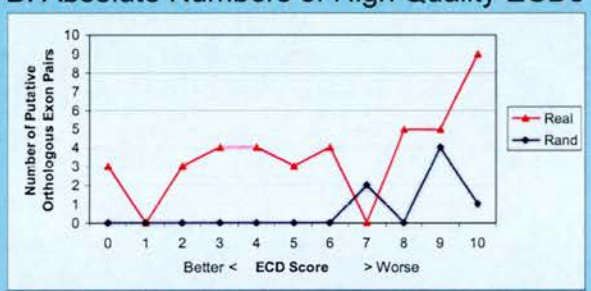
Section C shows the LOD scores generated from a comparison of these distributions. LOD scores are only given where 3 or more ECDs have been predicted in both the real and random distributions.

Figure 4.6. ECD Score Distributions and LOD Scores for Fish

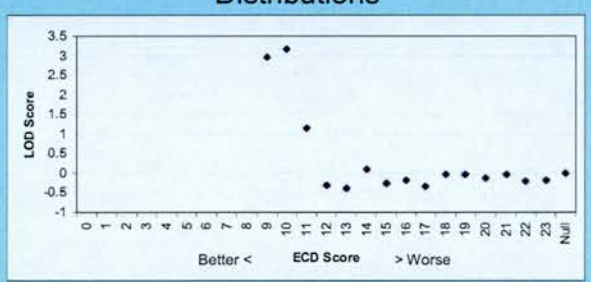
A. Cumulative Proportion of ECDs Below Score Threshold



B. Absolute Numbers of High Quality ECDs



C. LOD Scores Derived from these Distributions



Section A shows the cumulative distributions of all mouse:fish real and randomly generated ECD predictions. The green box indicates the region that is expanded in Section B.

Section B shows the absolute numbers of real and randomly generated ECD predictions with ECD scores less than or equal to ten. These are the highest quality ECD predictions.

Section C shows the LOD scores generated from a comparison of these distributions. LOD scores are only given where 3 or more ECDs have been predicted in both the real and random distributions.

an effort to further remove this source of error, second order polynomial trend-lines were fitted to each LOD score distribution, in an effort to smooth out the sample size errors. An example is shown in Figure 4.7 (p90).

The second order polynomials were calculated using an R^2 regression applied in Microsoft Excel. The values obtained were for the following equation, where x is the ECD score;

$$LOD = Ax^2 + Bx + C$$

These trend-lines were generated based only on the ECD scores for which the LOD score was greater than 0.2. All other ECD scores were not considered to be useful and so were given an automatic LOD score of zero. The ECS scores at which this occurs are also recorded in Table 4.2 (column D). The R^2 value provides an estimate of the goodness of fit.

Table 4.2. Polynomial Approximations of LOD Scores

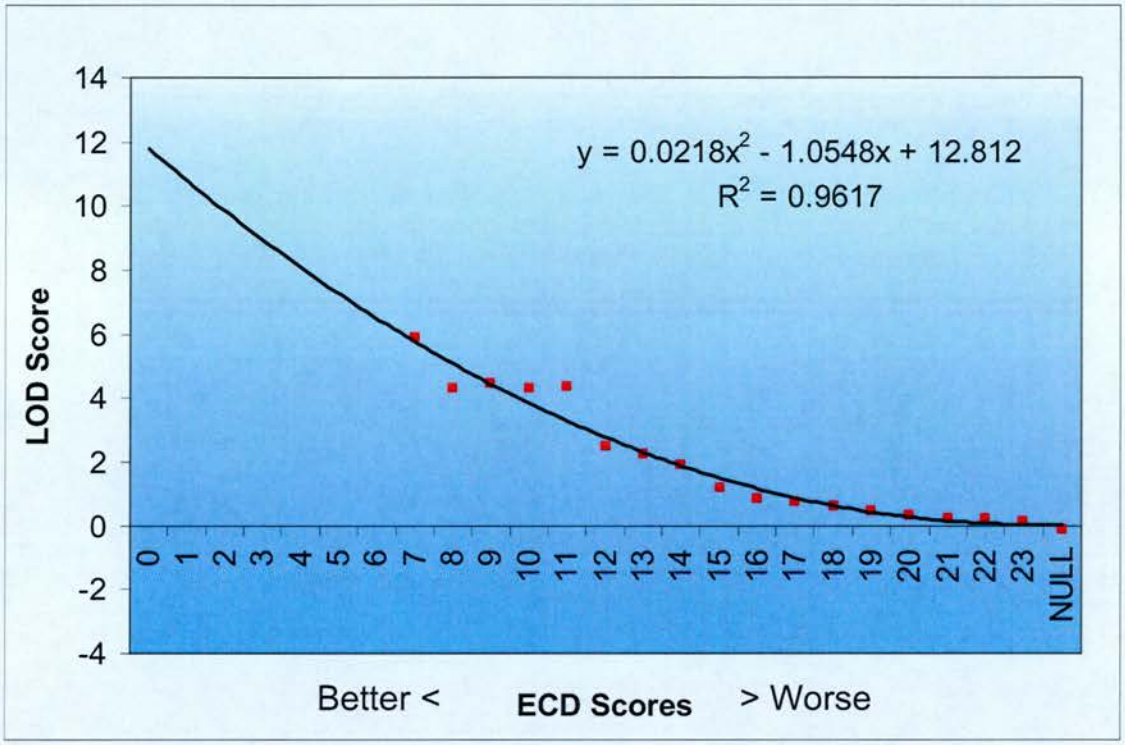
Species	A	B	C	D	R^2
Rat	0.0309	-0.8971	6.688	15	0.969
Human	0.0361	-1.3164	12.168	17	0.8789
Chicken	0.0218	-1.0548	12.812	20	0.9617
Fish	0.0317	-1.2515	11.457	13	0.7285
Fly	0.0138	-0.8931	14.102	25	0.6641

The values in this table were then used in conjunction with the above equation to calculate the LOD scores for each exon pair in each species pair. The best total LOD scores were then obtained for each mouse exon with at least one orthologous exon prediction. This was calculated by adding all the LOD scores for the top-scoring ECD prediction from each species pair for that mouse exon.

4.4.4 Annotation of the Putative Conserved ECDs

By ranking the resulting list of mouse exons, the best conserved ECD predictions were easily identified at the top of the list. These top ranking putative ECDs were then further investigated through a series of analyses.

Figure 4.7. Polynomial Approximation to the LOD Scores for Chicken



An example of a polynomial approximation to the LOD scores for chicken. LOD scores are derived from a comparison of the real and random distributions. These distributions are truncated and binned to ensure that each bin has at least 3 real and 3 random ECD predictions. The LOD scores derived from these bins are then smoothed (see Sections 4.4.3.1 & 4.4.3.2). Microsoft Excel is then used to approximate an polynomial trend-line to these values. This approximation accurately reflects the observed LOD scores ($R^2 = 0.9617$). The polynomial equation is given on the graph.

4.4.4.1 Identifying A-G Mismatches

To support the ECD predictions, all the Ensembl mouse exons were BLAST searched against all the publicly available transcribed mouse sequences in dbEST and GenBank (see Materials & Methods) and the available mouse genomic sequences (Options were –e 10e-10 –m 5). The resulting BLAST matches were scanned to identify any A-G mismatches that are seen between the exon and the expressed sequences, but not in the genomic sequences. Any mismatches meeting these criteria were recorded with the exon that they are found in.

4.4.4.2 Identifying Coding Exons

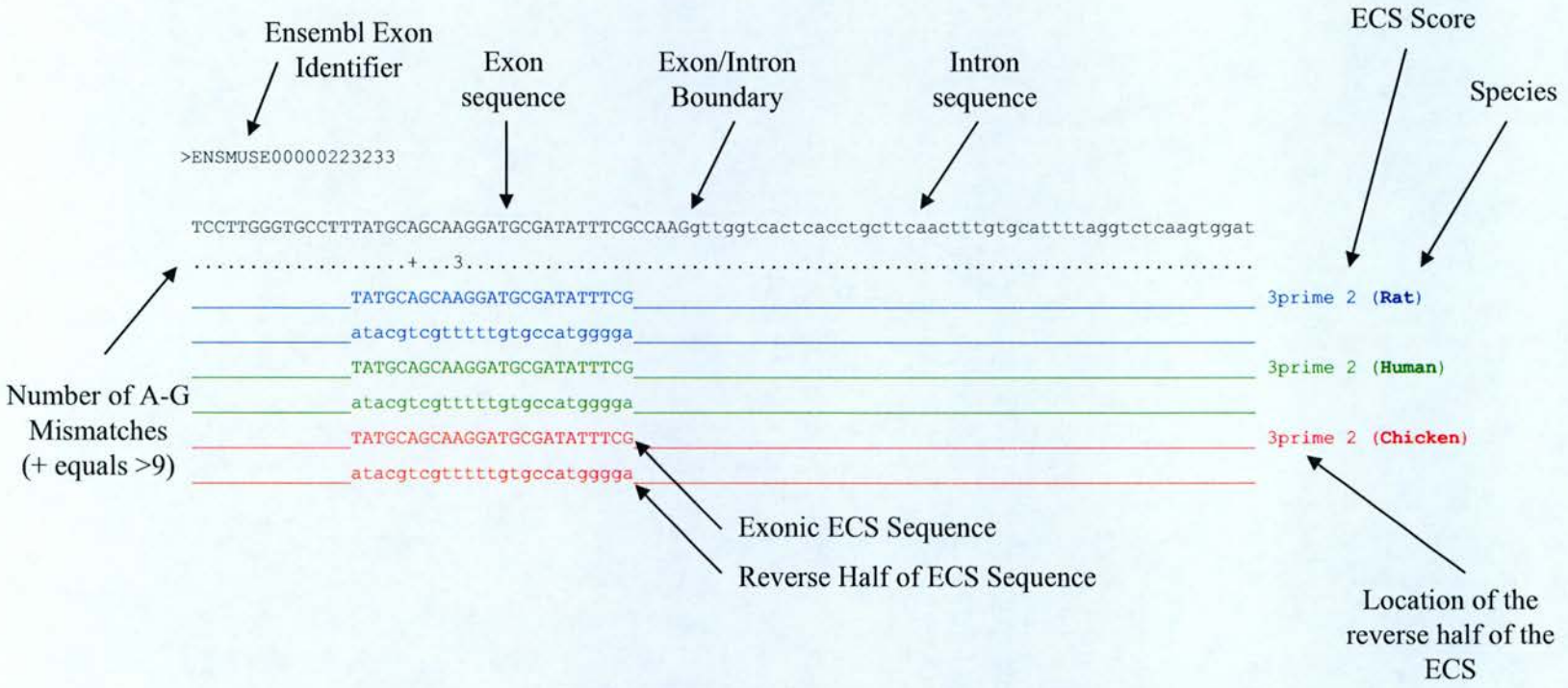
A table was obtained from EnsemblMart, which had the coordinates of any coding sequence in each mouse exon. This was used to annotate each exon as coding, non-coding or mixed.

4.4.4.3 Generating Alignment Reports

In order to fully analyse the remaining features of the putative ECDs an alignment report was generated. This included the exon and 50bp surrounding intronic sequence at either end. Both halves of each ECD scoring below a certain threshold were shown aligned to the un-gapped exon sequence. All A-G mismatches identified for this exon were also added to the alignment. See Figure 4.8 (p92) for an example. Initially, these reports were scanned to confirm that the top-scoring ECD from each pair of species were orthologous. Both halves of the ECD were required to overlap. Overlapping clusters of ECDs were manually recorded. Any A-G mismatches that overlapped these ECS clusters, or occurred nearby were also recorded.

The sequence qualities of the ECD clusters were also visually examined. This was primarily to search for low complexity sequences or repeats. In retrospect this was not a problem, as it appeared to be rare to find low complexity sequences or repeats conserved between distant species. Many of the known edited exons are edited at or

Figure 4.8. Sample Alignment Report



This is a sample taken from the GluR-B Q/R exon alignment. Only the best ECS predictions have been kept. Although this is a real example, it shows the best case scenario, where there are multiple ECSs that line up, with A-G mismatches overlapping them. There are often more potential ECSs identified in either the exonic or 5prime regions.

near the ends of the exon (see Figure 1.4 – p12). For this reason the distance to the nearest exon boundary was also measured from these alignments.

In order to facilitate easy viewing of these reports they have been re-formatted in HTML. These HTML reports include Javascript programs, which allow the viewer to change how the ECDs are displayed. These HTML reports are all available on the attached CD. This facility is explained in greater detail in Chapter 6.

4.4.5 The Finished Protocol

The protocol detailed here is the first method for identifying ECDs that are conserved between two or more species. The next section describes how successful this method is at identifying the known edited sites and their ECSs.

4.5 Results for Known Edited Exons

To determine the success of this protocol it needed to be tested on the known edited sites. Arguably, this introduced a degree of circularity as the protocol was based on the known sites. However, the protocol has only been based on loose generalisations of the known edited sites; i.e. they form conserved duplexes that have a minimum size of just over 25bp. One method that is often used in this situation is to base the model on half of the known sites and test it with the other half. Unfortunately, there are very few known recoding edited sites in vertebrates. However, the generalisations used here could have been obtained from any random selection of half the known sites. This suggests that the generalisations were broad enough that all the known edited sites could be used as validation.

The known vertebrate recoding edited sites are described in Figure 4.9 (p96). For the purposes of this project only the editing sites that do not occur in inverted Alu repeats, non-coding RNA or intronic RNA were used to test the protocol. Thus the focus was on edited sites that are known or assumed to affect function. They included the *GluR-B* (R/G & Q/R), *GluR-C* (R/G), *GluR-D* (R/G), *5HT$_{2C}$R*, *KCNA1*, *BC10*, *GluR-5*, *GluR-6*, *Cyfip2*, *Flna*, *EDNRB* and *Alpha 2,6-Sialyltransferase* sites. As the *IGFBP7* site was not experimentally confirmed, it has not been included in these analyses.

Although analysis was carried out on all these edited exons, they do not all have published data identifying their ECDs. Even fewer of these exons have experimental data supporting the published ECD predictions. Table 1.2 (p14) describes which sites have predictions, and what type of data supports them. Ten of the known editing site ECDs had at least some experimental validation in addition to the observation of editing in the ES. The different types of experimental validation are given in Table 1.2 (p14). These ten ECDs in Table 4.3 (p95) were used as positive controls to quantify the success of the method described in this chapter.

4.5.1 Finding the Known ECDs

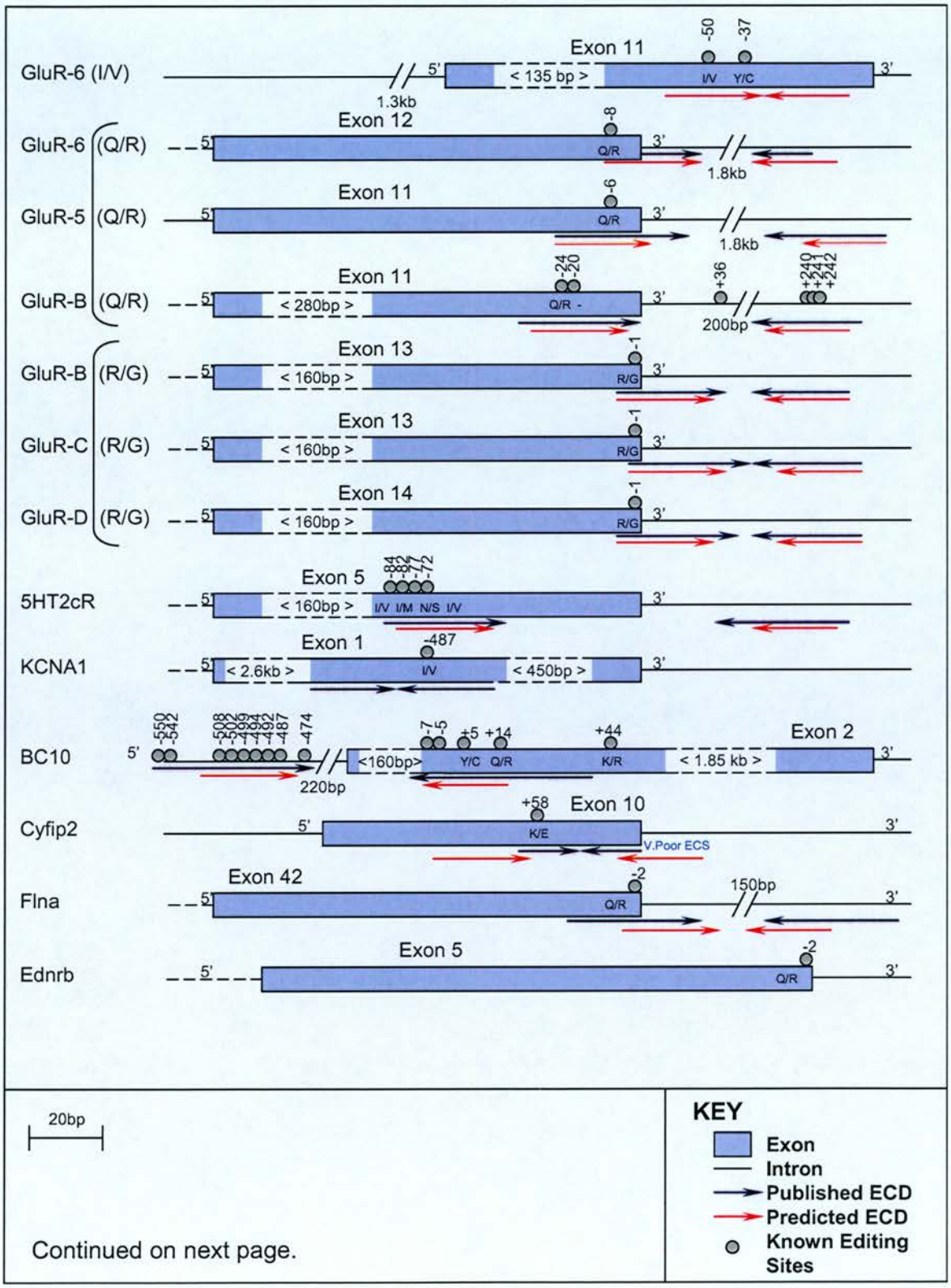
ECD predictions were made for all the known recoding edited sites, including those for which there are no experimentally validated ECS predictions. These are shown together with the published ECDs, where available, in Figure 4.9 (p96).

Table 4.3. Protocol Results for the Known Edited Sites

ECS Type	Known Edited Site	Top Rank in Mismatch Protocol[1]	Rank in Genome for ECS Protocol[2]	Best ECD Found with Mouse[3]														Comments[4]
				Rat			Human			Chicken			Danio			Fugu			
				Agree	Score	Where	Agree	Score	Where	Agree	Score	Where	Agree	Score	Where	Agree	Score	Where	
Experimentally Validated	GluR-B R/G	-	4th	Y	3	3P	Y	2	3P	Y	2	3P	Y	3	3P	Y	3	3P	Perfect match with known ECD
	GluR-C R/G	8th	1st	Y	2	3P	Y	1	3P	Y	2	3P	Y	2	3P	Y	2	3P	Perfect match with known ECD
	GluR-D R/G	9th	6th	No Orth.			Y	3	3P	Y	3	3P	Y	5	3P	Y	5	3P	Perfect match with known ECD
	GluR-B Q/R	6th	5th	Y	2	3P	Y	2	3P	Y	2	3P	-	-	-	-	-	-	Perfect match with known ECD
	GluR-5 Q/R	-	-	Y	6	3P	Y	10	3P	Y	10.5	3P	No Orth.			-	-	-	Perfect match with known ECD
	GluR-6 Q/R	-	42nd	Y	5	3P	Y	5	3P	Y	5	3P	No Orth.			-	-	-	Perfect match with known ECD. Some convincing exonic sites as well.
	5HT	1st	-	Y	3.5	3P	Y	4.5	3P	-	-	-	-	-	-	-	-	-	Perfect match with known ECD
	BC10	3rd	19th	Y	0	5P	Y	0	5P	-	-	-	-	-	-	-	-	-	Perfect match with known ECD
	ADAR2	-	-	Y	4	5P	Y	11.5	5P	-	-	-	-	-	-	Y	12	5P	Perfect match with known ECD
	KCNA	-	-	N	4.5	5P	N	6	5P	-	-	-	-	-	-	-	-	-	Best ECS is ~120 bp 5'prime of edited base.
Prediction Only	Flna	-	-	Y	4	3P	Y	5	3P	-	-	-	-	-	-	-	-	-	Perfect match with predicted ECD
	Cyfip2	-	-	O	9	3P	O	7.5	3P	-	-	-	-	-	-	-	-	-	Overlaps both halves of predicted ECD, but it is not identical. See Figure 4.10.
No prediction	GluR-6 I/V	-	154th	O	5	E	O	5	E	O	5	E	-	-	-	-	-	-	Best ECD overlaps Edited Site. See Figure 4.11
	Ednrb	-	-	N	7	E	N	9	E	-	-	-	-	-	-	-	-	-	No ECDs overlap the edited base.
	Alpha2,6-Sialyltransferase	-	-	N	7.5	3P	-	-	-	-	-	-	-	-	-	-	-	-	No ECDs overlap the edited base.

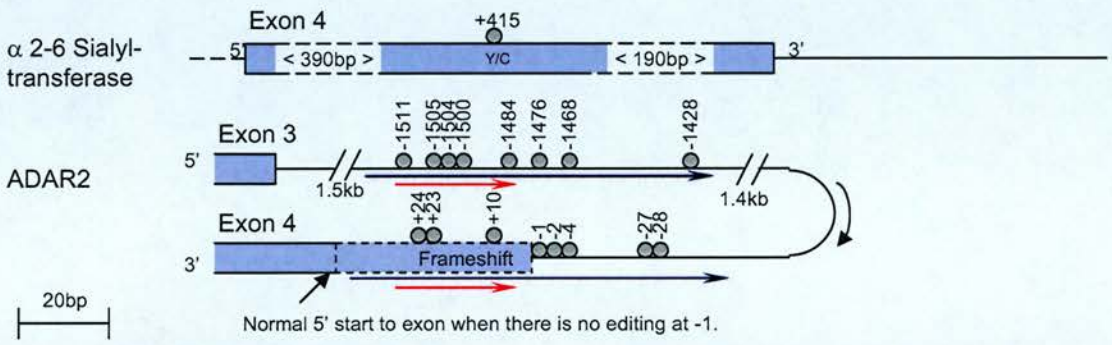
This table shows the results for the known protein recoding edited sites. These have been separated into three categories; those with experimentally validated ECD structures, those with published ECD predictions, but not validation, and those without any ECD predictions. **1.** Where the edited sites were found in the top 100 in the previous protocol (Chapter 3), their ranks have been shown. **2.** The rank of the known edited sites in the genome screen (only shown if in top 200). **3.** A breakdown of which species had good ECD scores (Score column), whether these predictions agree with the published ECDs (Agree column) and the location of the ECD (5'prime, 3'prime or Exonic)(Where column). Where orthologous exon predictions were not identified, these have been shown as 'No Orth'. **4.** Comments on these results. (These results are based on a core ECD size of 25bp, used to screen up to 2.5kb of intron in either direction).

Figure 4.9. The Known & Predicted ECDs for the Mammalian Protein Recoding Edited Sites



Continued on next page.

Figure 4.9. The Known & Predicted ECDs for the Mammalian Protein Recoding Edited Sites (Continued)



The 15 known protein recoding edited sites, showing the locations of the individual sites and their associated known ECSs. All transcripts are shown in the 5' to 3' direction and are drawn approximately to scale. The three glutamate receptor Q/R regions are grouped together due to their similarity, as are the three R/G regions. For the longer exons, gaps are shown with the intervening length given in the gap. Similarly, long intronic regions are denoted by two parallel slashes and an associated length. Where the protein is modified by editing, the amino acid change is indicated. The number above each site specifies the exact location of each site. Coordinates are relative to the nearest exon start/end. ADAR2 is alternatively spliced; editing at the –1 site creates an alternative splice site, resulting in the inclusion of 47bp and a frame shift. The resultant protein is not thought to be functional. The additional 47bp are indicated by the dashed line surrounding the initial segment of this exon. As well as giving an overview of the known sites, this figure demonstrates that mammalian recoding editing sites often occur in clusters and tend to be near the ends of exons. **The predicted ECD structures have been added as red arrows.**

KEY

- **Exon**
- **Intron**
- **Published ECD**
- **Predicted ECD**
- **Known Editing Sites**

These predictions are also available in HTML format on the attached CD. This protocol found the best predicted core ECD to overlap the known ECS and edited sites in nine out of the ten experimentally validated ECDs, with the exonic *KCNA1* site being the exception. This demonstrates a sensitivity of 0.9 and a high specificity given that only one false ECD was predicted. For each of these nine ECDs, the structure was found in at least three species.

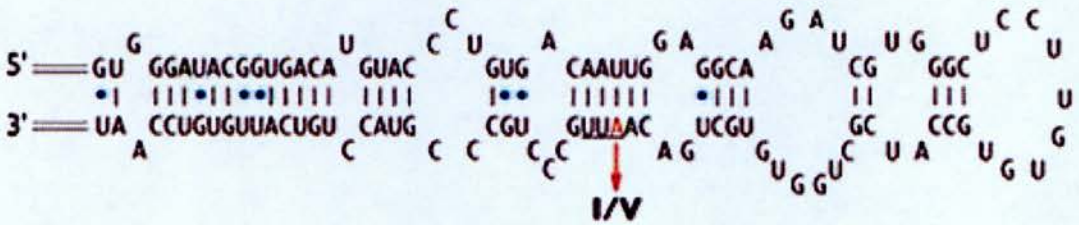
The most impressive observation from the results shown in Table 4.3 (p95) is that all three GluR R/G sites (*GluR-B,C & D*) are conserved exceptionally well in all species between mouse and both fish. This is in agreement with previous observations and represents conservation over an evolutionary distance of approximately 450 million years[86]. It is possible that this ECD is observable in more distant species, however that question has not been covered here.

Although they are not conserved across such a large evolutionary distance the *GluR-B* Q/R, *GluR-5* Q/R and *GluR-6* Q/R ECDs were all found in mouse, rat, human and chicken. This represents an evolutionary distance of approximately 310 million years[86]. The *GluR-B* Q/R site was of particular interest as this is proposed to be the single most important target of ADAR2 in the mouse[35].

The *5HT$_{2C}$R* and *BC10* sites were only seen in mouse, rat and human, however the quality of the *BC10* ECD is exceptional, with no gaps, bulges or mismatches in any of the three species. The *ADAR2* ECD was found in mouse, rat, human, and pufferfish in agreement with the published ECD[63]. These ECDs were not high quality, however. The chicken ECD prediction was so poor that it was not included in the table. The published ECD for this site is extremely long (~100bp), and the ECD we observed was only the best scoring part of it, although it did overlap the edited site. It is possible that this site makes up in size for its apparent lack of ECD quality.

The *KCNA1* ECD was the only positive control that was not identified by this protocol. This may be due to the fact that it forms a relatively imperfect duplex with lots of gaps and bulges, as shown in Figure 4.10 (p99). In this case the best predicted ECS formed between the 5 prime intron and the exon, approximately 120bp 5 prime of the edited nucleotide. Given that this ECS did not overlap the edited site or the known ECS, nor did it have particularly good scores, it was considered unlikely to be functional.

Figure 4.10. Published ECS Structure for KCNA1 Edited Site



This is the structure published by Bhalla *et al,* which has been validated using nucleotide substitutions and then functional rescue. The figure shown here is an adapted version of that used in the original paper. (Bhalla et al, Nat.Struct.Mol.Biol, 2004)

The success of this protocol at identifying these ECDs was likely to be biased by the fact that if these positive control sites did not have clear ECDs, then they would not have been published or tested. This bias was unavoidable and means that we could not rule out the presence of a class of ECDs with characteristics different from the positive controls.

4.5.2 Novel Predictions for Known Edited Sites

The remaining ECD predictions for the known edited sites represented novel data. There were five edited sites without experimentally defined ECSs, including the *GluR-6* I/V, *Cyfip2, Flna, Ednrb* and *Alpha 2,6 sialytransferase* sites. Two of these sites had published ECS structures predicted by MFOLD. Of these, the Flna ECD prediction agreed with the MFOLD structure, while the *Cyfip2* ECD prediction overlapped the MFOLD structure, but did not exactly agree. This situation is shown in Figure 4.11 (p101), which shows the relative locations of the two ECD predictions in addition to a third prediction using RNAFold on the displayed sequence. The RNAFold prediction supported the new prediction, however the results from these programs can be unreliable[175].

The *GluR-6* I/V best ECD prediction also overlapped the known edited nucleotides, despite the exon being fairly long (224bp). As shown in Figure 4.12 (p101), both halves of this ECD occurred within the exon. Given that an identical ECS formed in mouse, rat, human and chicken, it seemed that this is very likely to be a functional ECS, especially given the relatively high quality of the ECD (i.e. low score). However, on closer analysis there was a second putative ECD, which again overlapped the edited base, but formed with the 3 prime intron. This ECD scored slightly better, but was only observable in mouse, rat and human. It is possible that both of these ECSs are functional, however, determining this would require experimental validation.

Neither of the ECD predictions for the remaining two edited sites overlapped the edited bases, nor were their scores particularly convincing (*Ednrb & Alpha 2,6 sialyltransferase*).

Figure 4.11. Sequence Conservation & Predicted Structures for the Cyfip2 Site

```
Edit                          *
Mouse AUAGAGCUGGCCAGAUACAUUAAGACCAGUGCUCACUAUGAAGAGAACAAGUCCAAgugagugccugccguaaugucucuuggcu
Rat   AUAGAGCUGGCCAGAUACAUUAAGACCAGUGCUCACUAUGAAGAGGACAAAUCCAAgugagugccugccuuacugucucuuggcu
Human AUAGAGCUGGCCAGAUACAUUAAGACCAGUGCUCACUAUGAAGAGAACAAGUCCAAgugagugccugccguaaugucucucggcu
Gall  AUAGAACUGGCCAGAUACAUUAAGACCAGUGCUCACUAUGAGGAGAACAAAUCCAAgugagugccugccauaaugucucuuggcu
Cons. ***** ********************************** *** **** ****************** ** ******* ****
LevECS                  >>>>>>>>>>>>>>>>>>          <<<<<<<<<<<<<<<<<
MyECS               >>>>>>>>>>>>>>>>>>>>>>>>>>                      <<<<<<<<<<<<<<<<<<<<<<<<<
RFold .......>>>>>>>.>>>>>>....>>>>>>>>>>>.>>.............<<<<<<<<<<.<<<...<<<<<<.<<.<<<<<
```

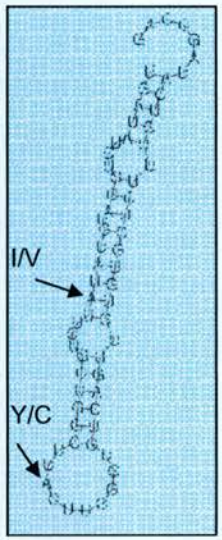
A sequence alignment and predicted ECS structures for the Cyfip2 edited site. The edited site is shown on top of the sequence. There are three ECS predictions shown, including the prediction published by Levanon et al (LevECS), the prediction from this protocol(MyECS) and a prediction from the RNA Fold algorithm, given the above mouse sequence (Rfold). The three ECSs do not agree entirely, but they all overlap.

Figure 4.12. Sequence Conservation & Predicted Structures for the GluR-6 I/V Site

A

```
Edit                  I/V          Y/C
Edit                   *            *
Mus   UGAUAUCUGGAUGUAUAUUCUGCUGGCUUACUUGGGUGUCAGUUGUGUGCUCUUUGUCAUAGCCAG
Rat   UGAUAUCUGGAUGUAUAUUCUGCUGGCUUACUUGGGUGUCAGUUGUGUGCUCUUUGUCAUAGCCAG
Hum   UGAUAUCUGGAUGUAUAUUCUGCUGGCUUACUUGGGUGUCAGUUGUGUGCUCUUUGUCAUAGCCAG
Gal   UGAUAUCUGGAUGUAUAUUCUGCUGGCUUACUUGGGUGUCAGUUGUGUGCUCUUUGUCAUAGCCAG
Cons. ************************************************************************
MyECS              >>>>>>>>>>>>>>>>>>> <<<<<<<<<<<<<<<<<<
Rfold >>>>>...>>>.>>>>>>...>>>>>>.........<<<<<<.<<<<<<<<<.<<<<<.......
```

B



Section A: A sequence alignment and predicted ECS structures for the GluR-6 I/V edited site. The edited sites are shown on top of the sequence. The ECS prediction supplied by this work appears similar, although not identical to the RNAFold prediction. **Section B**: The structure predicted by the RNA Fold algorithm.

Given that convincing ECDs were not found for all of the known edited sites in all the species, an additional, more sensitive analysis was carried out. This screen looked further away, up to 10kb from the exon ends, including any exonic sequence in these flanking regions. This screen was also more thorough in comparing the predicted ECDs from each species, by taking the top 40 putative ECDs from each species rather than the top 20 (see Materials & Methods). No convincing ECDs were identified in addition to those already described. This suggested that where we have failed to identify conserved ECSs, either they do not exist, or they are of poor quality and hence difficult to detect.

These analyses clearly show that the protocol described here was capable of identifying known ECS structures over a range of evolutionary distances from 40 to 450 million years of divergence[86] and that it was sensitive to both known editing sites and their ECSs. It would also appear to provide ECD predictions of comparable utility to those from more sophisticated, and more computationally demanding RNA secondary structure prediction programs. To assess the specificity of this protocol and to identify novel edited candidates, a full analysis of every exon in the mouse genome was performed. Randomised controls were also generated in an effort to provide meaningful statistical analysis. The results of these genome screens are given in the following section.

4.6 A Mouse Genome ECS Screen

As described in the Materials & Methods, LOD scores were derived for the top-scoring ECD in each exon, using both the correct introns and randomly selected quadruplets of introns. Due to the nature of the randomisations, these results only applied where the ECS half of the ECD is found in a flanking intron. Combined LOD scores were then obtained by adding the LOD scores from each species pair together for a given mouse exon. In theory these LOD scores could have been converted into probabilities, but due to issues with the independence of the individual LOD scores, this was considered improper. Instead, the combined LOD score was used only as a method to rank the mouse exons in an attempt to identify the best conserved ECDs.

A summary of the top 50 exons is given in Table 4.4 (p104). Alignments for the 30 top-scoring vertebrate exons are included in Appendix 1, together with additional information on the gene functions, and the structure and location of the ECD predictions. The attached CD includes the 50 top-scoring exons in HTML format, as well as a Microsoft Excel table of the top 1,000 exons.

4.6.1 Performance of the Known Editing Sites

The best scoring end of the distribution of combined LOD scores is shown in Figure 4.13 (p107). In contrast to the ECD scores, the best LOD scores are the highest ones. The positions and ranks of the high scoring known edited sites are shown in this figure and Table 4.5 (p108), respectively. This shows that there was a strong cluster of known editing sites at the high scoring end of the distribution. The ranks were out of a total of 221,626 mouse exons.

Once again the results for the three GluR R/G sites were very impressive coming 1st, 4th and 6th out of all mouse exons. *BC10* ranked impressively given that it is only found in the mouse:rat and mouse:human comparisons (19th). This was due to the exceptionally good ECD scores that each analysis gave in these species. The results for the two *GluR-6* sites and the *Flna* site were also encouragingly high.

The remainder of the known edited exons had less impressive combined LOD scores. This showed that, although this protocol was effective for identifying ECDs for some

Table 4.4. ECS Predictions for the 50 Top-Scoring Exons in Vertebrates

Rank	Mouse Exon	Mouse Gene	LOD Score	ECDs Overlap?	Known Edited Site/Gene or Class	Gene Description from Ensembl Mouse
1st	ENSMUSE00000478281	ENSMUSG00000001986	35.8	RHGDF	GluR-C R/G	Glutamate receptor, ionotropic, AMPA3 (alpha 3); cAMP-dependent Rap1 guanine-nucleotide exchange factor.
2nd	ENSMUSE00000477286	ENSMUSG00000033981	34.6	RHGF	GluR-B	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
3rd	ENSMUSE00000490396	ENSMUSG00000001986	34.6	RHGDF	GluR-C	glutamate receptor, ionotropic, AMPA3 (alpha 3).
4th	ENSMUSE00000223221	ENSMUSG00000033981	32.7	RHGDF	GluR-B R/G	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
5th	ENSMUSE00000223233	ENSMUSG00000033981	25.5	RHG	GlurR-B Q/R	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
6th	ENSMUSE00000519849	ENSMUSG00000025892	24.4	HGDF	GluR-D R/G	Glutamate receptor 4 precursor (GluR-4) (GluR4) (GluR-D) (Glutamate receptor ionotropic, AMPA 4).
7th	ENSMUSE00000105648	ENSMUSG00000020524	20.5	HG	GluR-A	Glutamate receptor 1 precursor (GluR-1) (GluR-A) (GluR-K1) (Glutamate receptor ionotropic, AMPA 1).
8th	ENSMUSE00000515006	ENSMUSG00000025892	20.5	HG	GluR-D	Glutamate receptor 4 precursor (GluR-4) (GluR4) (GluR-D) (Glutamate receptor ionotropic, AMPA 4).
9th	ENSMUSE00000479443	ENSMUSG00000007850	20.1	RH	hnRNP	Heterogeneous nuclear ribonucleoprotein H (hnRNP H).
10th	ENSMUSE00000113629	ENSMUSG00000047022	20.0	RHG	-	Mirror-image polydactyly gene 1 protein homolog.
11th	ENSMUSE00000228059	ENSMUSG00000033569	20.0	RHG	-	Brain-specific angiogenesis inhibitor 3 precursor.
12th	ENSMUSE00000177912	ENSMUSG00000028289	20.0	RHG	-	Ephrin type-A receptor 7 precursor (EC 2.7.1.112) (Tyrosine-protein kinase receptor EHK-3) (EPH homology kinase-3)
13th	ENSMUSE00000113867	ENSMUSG00000061603	20.0	HG	-	NA
14th	ENSMUSE00000186519	ENSMUSG00000029169	20.0	RHG	Splicing	Putative pre-mRNA splicing factor RNA helicase (DEAH box protein 15).
15th	ENSMUSE00000496440	ENSMUSG00000029563	20.0	RHG	-	Forkhead box protein P2.
16th	ENSMUSE00000375270	ENSMUSG00000028546	19.3	RH	Splicing	ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD) (Hu-antigen D).
17th	ENSMUSE00000466378	ENSMUSG00000025255	19.2	HG	-	zinc finger homeodomain 4.
18th	ENSMUSE00000476281	ENSMUSG00000025255	19.2	HG	-	zinc finger homeodomain 4 - Exon is essentially the same as the one above. ECSs are identical.
19th	ENSMUSE00000472446	ENSMUSG00000057453	18.9	RH	BC10	Bladder cancer-associated protein (Bladder cancer 10 kDa protein) (Bc10).

Table 4.4. ECS Predictions for the 50 Top-Scoring Exons in Vertebrates

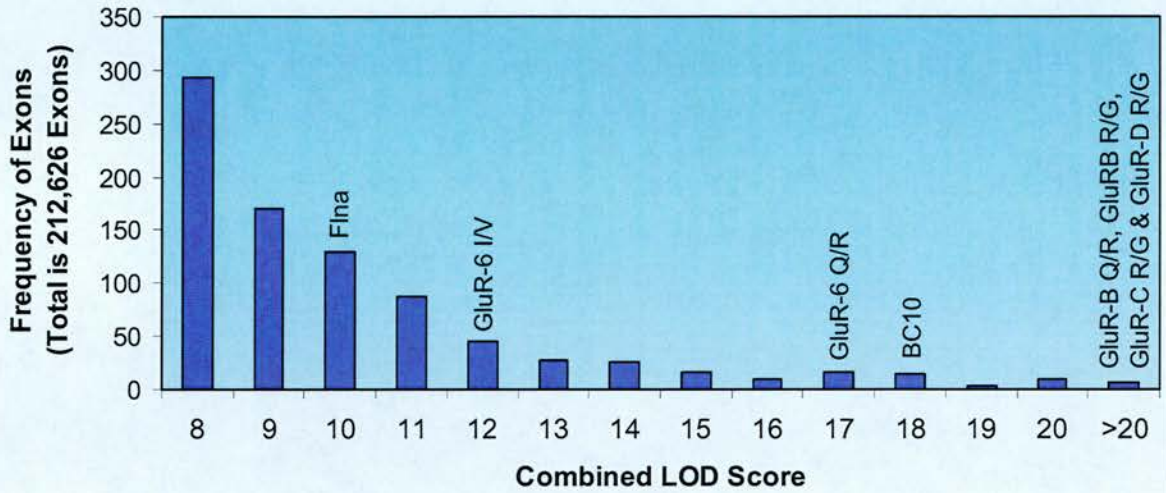
Rank	Mouse Exon	Mouse Gene	LOD Score	ECDs Overlap?	Known Edited Site/Gene or Class	Gene Description from Ensembl Mouse
20th	ENSMUSE00000165486	ENSMUSG00000027018	18.7	RHG	Chromatin	histone aminotransferase 1; histidine aminotransferase 1.
21st	ENSMUSE00000327165	ENSMUSG00000037369	18.7	RHG	-	Ubiquitously transcribed X chromosome tetratricopeptide repeat protein (Ubiquitously transcribed TPR protein on the X)
22nd	ENSMUSE00000195189	ENSMUSG00000030096	18.5	RHG	Synaptic	Sodium- and chloride-dependent taurine and beta-alanine transporter.
23rd	ENSMUSE00000207263	ENSMUSG00000006678	18.5	HG	PolyA	DNA polymerase alpha catalytic subunit (EC 2.7.7.7).
24th	ENSMUSE00000395043	ENSMUSG00000031284	18.3	RHG	-	Serine/threonine-protein kinase PAK 3 (EC 2.7.1.37) (p21-activated kinase 3) (PAK-3) (Beta-PAK) (CDC42/RAC effector
25th	ENSMUSE00000463801	ENSMUSG00000033565	18.2	RHG	-	RNA binding motif protein 9; fox-1 homolog (C. elegans); hexaribonucleotide binding protein 2; Fyn-binding molecule
26th	ENSMUSE00000099319	ENSMUSG00000019947	18.2	RHG	-	AT-rich interactive domain-containing protein 5B (ARID domain- containing protein 5B) (Mrf1-like)
27th	ENSMUSE00000153249	ENSMUSG00000025789	18.2	RG	Neurogenesis	Alpha-2,8-sialyltransferase 8B (EC 2.4.99.-) (ST8Sia II) (Sialyltransferase X) (STX) (Polysialic acid synthase).
28th	ENSMUSE00000469222	ENSMUSG00000033981	18.2	RHG	GluR-B	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
29th	ENSMUSE00000487901	ENSMUSG00000033981	18.2	RHG	GluR-B	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2). Exon is essentially same as above.
30th	ENSMUSE00000178321	ENSMUSG00000061455	18.1	RD	Synaptic	NA - Syntaxin variant
31st	ENSMUSE00000504265	ENSMUSG00000020385	18.1	RHG	Splicing	Dual specificity protein kinase CLK4 (EC 2.7.1.37) (EC 2.7.1.112) (CDC like kinase 4).
32nd	ENSMUSE00000491207	ENSMUSG00000037581	18.0	RHG	-	Ribosomal protein S6 kinase polypeptide 1
33rd	ENSMUSE00000234215	ENSMUSG00000033565	18.0	RH	-	RNA binding motif protein 9; fox-1 homolog (C. elegans); hexaribonucleotide binding protein 2; Fyn-binding molecule
34th	ENSMUSE00000426424	ENSMUSG00000028546	17.9	RH	Splicing	ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD) (Hu-antigen D).
35th	ENSMUSE00000342145	ENSMUSG00000028546	17.9	RH	Splicing	ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD) (Hu-antigen D).
36th	ENSMUSE00000155429	ENSMUSG00000060679	17.9	RH	hnRNP	mitochondrial ribosomal protein S9; muscle protein 637.
37th	ENSMUSE00000166139	ENSMUSG00000027108	17.7	RHG	-	Putative GTP-binding protein PTD004 homolog.
38th	ENSMUSE00000163299	ENSMUSG00000026833	17.7	HF	Neurogenesis	Noelin precursor (Neuronal olfactomedin-related ER localized protein) (Olfactomedin 1) (Pancortin).

Table 4.4. ECS Predictions for the 50 Top-Scoring Exons in Vertebrates

Rank	Mouse Exon	Mouse Gene	LOD Score	ECDs Overlap?	Known Edited Site/Gene or Class	Gene Description from Ensembl Mouse
39th	ENSMUSE00000462233	ENSMUSG00000025255	17.6	RHG	-	zinc finger homeodomain 4.
40th	ENSMUSE00000138504	ENSMUSG00000024120	17.5	RHG	-	leucine-rich PPR motif-containing protein; leucine rich protein LRP130.
41st	ENSMUSE00000165494	ENSMUSG00000027018	17.5	RHG	Chromatin	histone aminotransferase 1; histidine aminotransferase 1.
42nd	ENSMUSE00000495901	ENSMUSG00000056073	17.5	RHG	GluR-6 Q/R	Glutamate receptor, ionotropic kainate 2 precursor (Glutamate receptor 6) (GluR-6) (GluR6) (Glutamate receptor
43rd	ENSMUSE00000166138	ENSMUSG00000027108	17.5	RHG	-	Putative GTP-binding protein PTD004 homolog.
44th	ENSMUSE00000224458	ENSMUSG00000027168	17.5	RH	-	Paired box protein Pax-6 (Oculorhombin).
45th	ENSMUSE00000287940	ENSMUSG00000031302	17.3	RH	Synaptic	Neuroligin 3 precursor (Gliotactin homolog).
46th	ENSMUSE00000191872	ENSMUSG00000029705	17.3	RH	-	CCAAT displacement protein (CDP) (Cut-like 1) (Homeobox protein Cux) (Fragment).
47th	ENSMUSE00000110929	ENSMUSG00000014349	17.1	RH	-	NA - Involved in Ubiquition conjugation
48th	ENSMUSE00000163467	ENSMUSG00000060817	17.1	RG	-	NA
49th	ENSMUSE00000438324	ENSMUSG00000052049	17.1	-	Neurogenesis	SLIT and NTRK-like protein 1 precursor.
50th	ENSMUSE00000312703	ENSMUSG00000025224	16.8	RHG	-	Golgi-specific brefeldin A-resistance guanine nucleotide exchange factor 1.

The 50 top-scoring exons in the screen of six vertebrate species. Each exon is listed with its Ensembl exon and gene identifiers and the gene description. A high combined LOD score indicates a convincing conserved ECD structure. The 'ECD Overlap' column shows which species have orthologous ECD predictions. Both the ES and the ECS must overlap to be considered orthologous. Where the exon is a positive control the row has been shaded a darker blue. Novel candidate exons are light blue. Where the exons are from genes that are known to contain recoding edited sites, the names of these genes are annotated. In the same column, if a gene fits into one of the following functional classes it has been annotated as such: Splicing machinery, Synaptic machinery, hnRNPs, Chromatin regulation or Neurogenesis.

Figure 4.13. Distribution of the Top 1,000 Combined LOD Scores



This figure shows the high scoring end of the frequency distribution of combined LOD scores for all mouse exons. Higher LOD scores indicate better ECD predictions. The known edited sites that occurred in this part of the distribution are indicated above their respective LOD scores. 99.5% of all mouse exons have LOD scores lower than 8.

of the known edited exons, it was not powerful enough to detect all the known edited exons in a genome screen with high specificity. It was, however, extremely specific for some of the better ECDs.

Table 4.5 Performance of the Known Edited Sites

Known Site	Rank	Percentage Scoring Higher
GluR-C R/G	1st	0%
GluR-B R/G	4th	0.0014%
GluR-B Q/R	5th	0.0018%
GluR-D R/G	6th	0.0027%
BC10	19th	0.0086%
GluR-6 Q/R	42nd	0.0194%
GluR-6 I/V	153rd	0.0690%
Flna	382nd	0.1724%

In an ideal situation, precise calculations of true sensitivity and specificity should have been carried out. However, these calculations require both a set of ECDs from known edited exons and a set of exons that are known not to be edited. While the first set exists, albeit with a small sample size, the latter does not exist. It would be very difficult to prove an exon is not edited in any tissue, life stage or environmental situation. Instead we can falsely assume that every exon that is not a positive control is a negative control.

Using a combined LOD score threshold of 24, we obtained four of the 10 exons with known and validated ECSs and only two other exons (both of which were additional exons from the known edited genes). This represented an artificially high specificity of almost 100% and a relatively low sensitivity of 40% (based on the following standard equations).

$$Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP}$$

Using a looser threshold of 17.5, we obtained six of the 10 exons with known ECDs and only 43 other exons, of which five were additional exons from the known edited genes. Again this represented an artificially high specificity of almost 100%, but with a slightly improved sensitivity of 60%.

It has now been clearly established that this protocol was able to specifically identify many of the known editing sites. However, given that this protocol was partly based on observations from these sites, these results may have been artificially impressive. The real test of the protocol was in the identification of novel editing sites.

4.6.2 Candidate Editing Sites

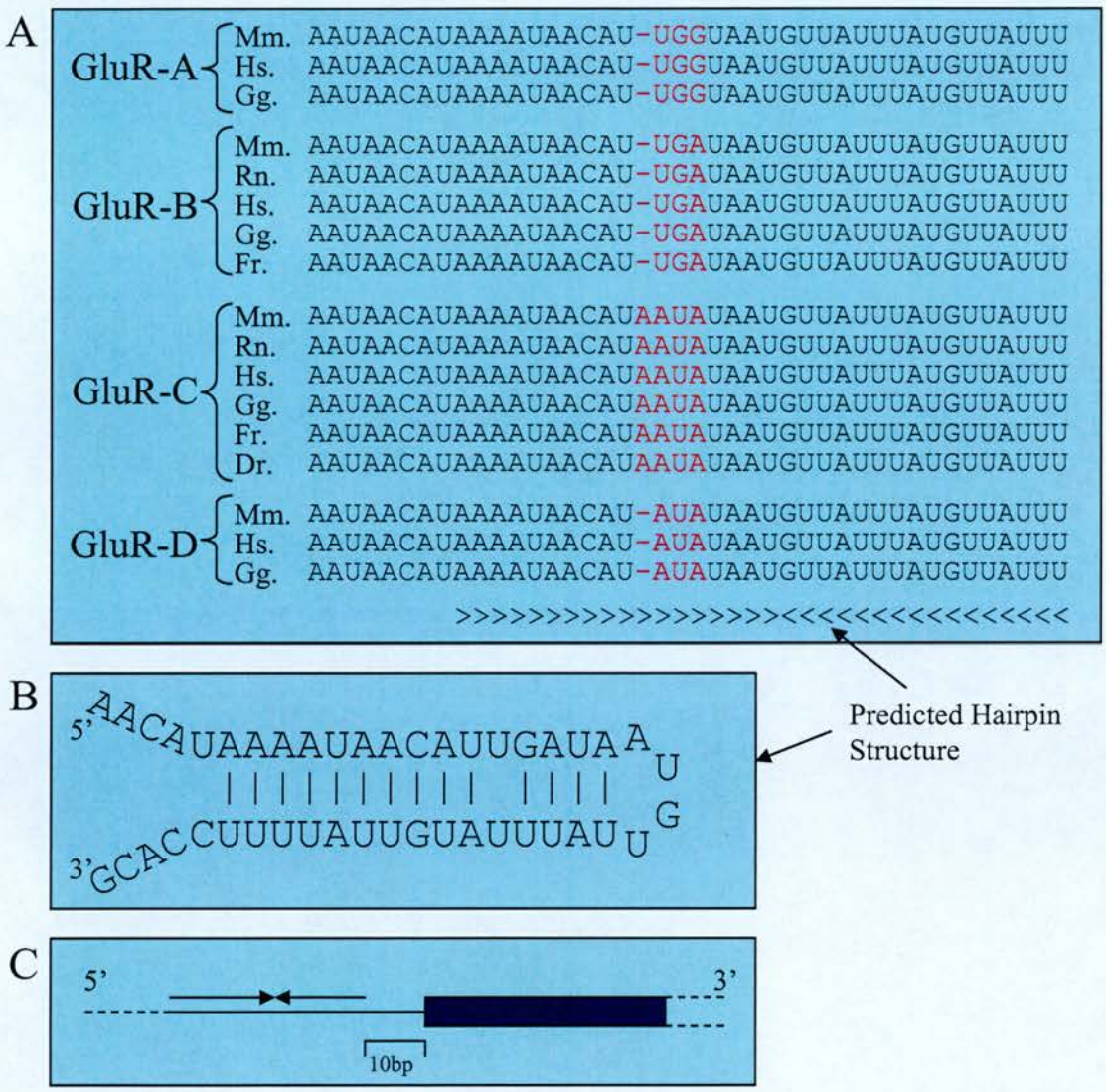
The 50 top scoring exons in the mouse genome, shown in Table 4.4 (p104), included six known edited exons, five additional exons from glutamate receptors, three duplicate results and 36 other exons. The three duplicate pairs of exons were due to alternative splicing at the 5 prime or 3 prime ends to the exon resulting in overlapping exons having different Ensembl exon IDs. This resulted in the protocol treating them as separate exons, both of which scored identically. The ranks of these exons were $17^{th}/18^{th}$, $28^{th}/29^{th}$ and $34^{th}/35^{th}$ (from *ZF Homeodomain 4*, *GluR-B* and the *ELAV* gene, respectively).

4.6.3 Another Conserved Glutamate Receptor ECD?

It is very interesting that five of the novel candidate exons were from glutamate receptors, which are clearly good targets for editing in vertebrates. Closer examination of these sites showed that the first four, which rank 2^{nd}, 3^{rd}, 7^{th} and 8^{th}, were homologous exons from *GluR-B, GluR-C, GluR-A* and *GluR-D* respectively. In each case the ECD predicted was practically identical for each species observed. This is shown in Figure 4.14 (p110). Interestingly, both halves of this ECD lie adjacent in the sequence, forming a tight predicted hairpin. The whole structure resides in the 50bp 5 prime of the exon start. This exon is termed the flip exon and is alternatively spliced (see Introduction).

The degree of sequence conservation shown here is remarkable, especially given that this is intronic sequence and that the conservation is observable not only across a vast evolutionary range (mouse to fish), but also between each of the four genes. Figure 4.14 (p110) also shows the predicted hairpin structure, which although short, is of good quality. It is surprising that this ECS should have been so well conserved yet not overlap any protein coding sequence. This suggested that the ECD might function in the regulation of splicing. To investigate this, potential splice branch site

Figure 4.14. Intronic Sequence Conservation Between Four Glutamate Receptors.



Section A: Sequence alignment of the GluR-A, GluR-B, GluR-C & GluR-D in all each species that this site was observed in. The sequences in red indicate the ONLY bases that vary in this alignment, between species or genes.

Section B: Putative ECS structure for the mouse GluR-B ECS. This structure is also indicated in section A.

Section C: Location of the ECS halves in relation to the 5 prime end of the exon.

sequences were identified using a branch site consensus model[176], as shown in Figure 4.15 (p112). Branch sites occur between 18bp and 40bp upstream of the exon start, meaning all the potential branch sites were covered by the predicted ECD structure. In particular there was one potential branch site that is seen in all the sequences, suggesting that this is the functional one. The essential part of a branch site is the adenosine in the sixth position. Given that this branch site occurs in an RNA duplex it is tempting to think that this adenosine could be edited, thus removing the branch site and resulting in an alternative splice form. Figure 4.16 (p113) shows that there were three observable splice forms, although others may exist. Splice form C results in a frame-shift and has not been published, and is of low abundance, so it is likely to be an experimental artefact. Examination of the literature showed that the other two splice forms are well documented as the 'flip' and 'flop' isoforms (splice forms B & A, respectively). These splice variants have a strong effect on the gating properties of the AMPA receptors[177].

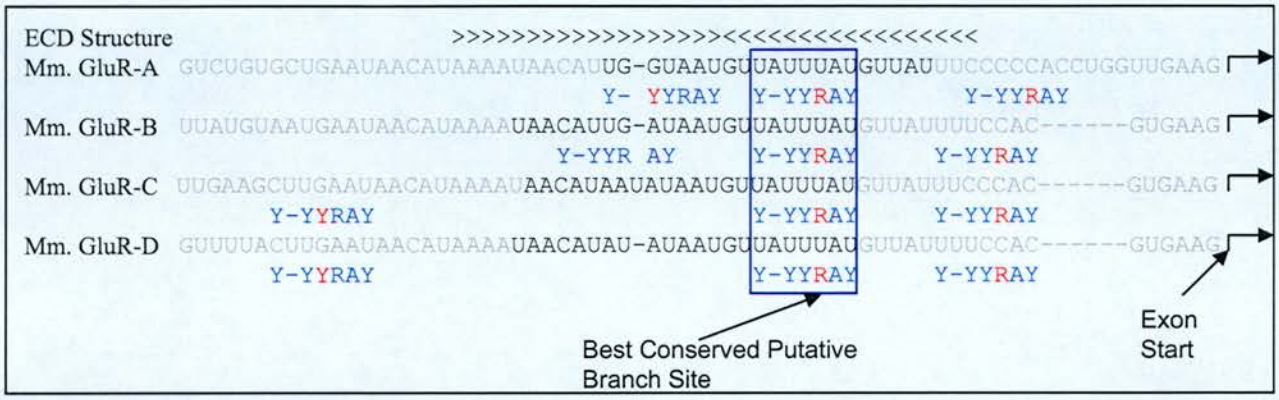
This hypothesis is further supported by the finding that when editing of the *GluR-D* R/G is low, the 'flip' isoform is more prevalent, suggesting that, in at least some cell types, there is interplay between the splicing of this exon and editing[32]. In conclusion, these are exceptionally well conserved intronic ECDs that cover, and could remove, a vital branch site for the 'flip' isoform. These predictions are being experimentally tested to determine if there is editing at the predicted branch site. This work is still in progress.

The remaining novel candidate exon found in a glutamate receptor was not related to the other four glutamate sites discussed above. There was not sufficient evidence for this exon to consider testing it. Figure 4.17 (p114) shows the locations of the known edited sites and the 'flip/flop' exon in relation to the protein structure of a typical ionotropic glutamate receptor[30].

4.6.4 Novel Edited Sites in Other Genes

There are 36 remaining novel candidate exons in the top 50 exons. The fact that all these exons contained protein-coding sequence was encouraging, however, these tend to be the best conserved sequences, so it was not particularly surprising to see this data set enriched for them. Due to the nature of the protocol it was not possible to pinpoint the amino acids that might be altered by editing. Instead, a region of the

Figure 4.15. Putative Splice Branch Sites in the Conserved Intronic Regions.



Splice Branch Site Consensus

$Py_{80}NPy_{80}Py_{87}Pu_{75}A_{100}Py_{95}$

Y = Py = Pyrimidine (C/U)

R = Pu = Purine (A/G)

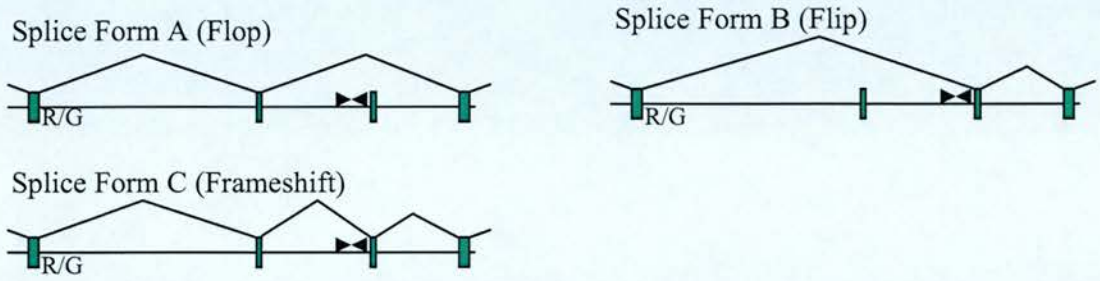
Red text indicates that the base does not fit the consensus.

Grey text indicates that the base is not within the normal region for branch sites (which are normally 18-40bp from the 3 prime exon).

Putative splicing branch points for four Glutamate receptor exons (GluR-A,B,C&D) in relation to the predicted conserved ECD structures. The ECD structure is shown aligned to the four intronic sequences that contain it. Branch sites normally occur between 18bp and 40bp upstream of the 3 prime exon. This region is indicated by the nucleotides in black. A branch site consensus is shown in the bottom left hand panel. The numbers indicate the proportion of branch sites with a purine/pyrimidine at each position. This consensus was used to identify potential branch sites in the appropriate regions. There is one putative branch site that exists in all four introns and within the allowed regions. This is likely to be the functional branch site.

Figure 4.16. Splice Isoforms of the Four Glutamate Receptors & cDNA Evidence

A



B

GluR-A Species	Splice Form A	Splice Form B	Splice Form C
Mouse	BC056397	X57497	NA
Rat	X17184	M38060	NA
Human	M81886	AB209094	NA
Chicken	U59704	X89510	NA
Zebrafish	AF525741	AF525744	NA

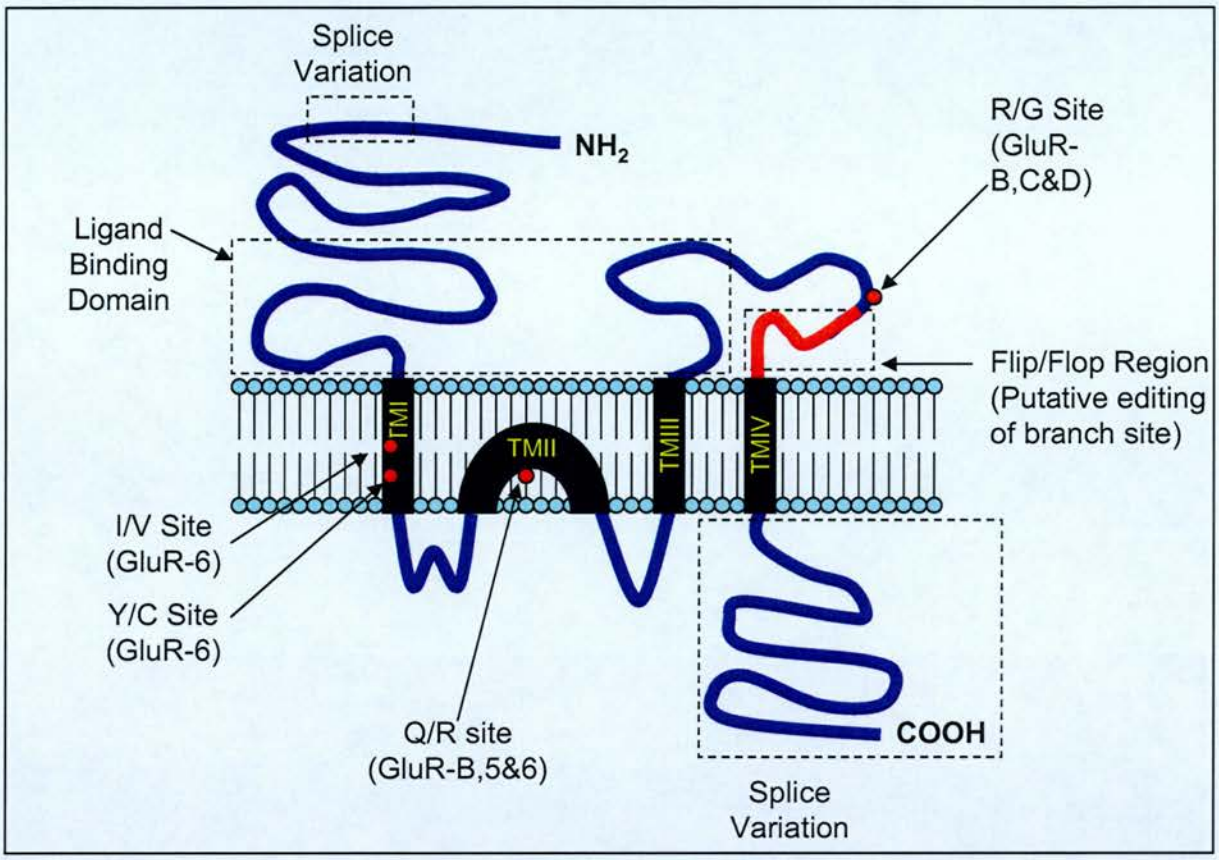
GluR-B Species	Splice Form A	Splice Form B	Splice Form C
Mouse	X57498	AB111957	BC066193
Rat	M85035	M38061	NA
Human	BC010574	AB209567	AV747836
Chicken	U59706	X89508	NA
Zebrafish	AF525743	AF525744	NA

GluR-C Species	Splice Form A	Splice Form B	Splice Form C
Mouse	NA	AB022342	BC076584
Rat	M85036	X54656	NA
Human	U10302	U10301	NA
Chicken	X89509	U59709	NA
Zebrafish	AF525746	AF525745	NA

GluR-D Species	Splice Form A	Splice Form B	Splice Form C
Mouse	AB022913	AK031568	NA
Rat	M36421	M85037	NA
Human	NA	U16129	NA
Chicken	X89507	U59707	NA
Zebrafish	AF525748	AF525747	NA

Section A shows the three mRNA isoforms that were observable in publicly available cDNAs and ESTs. Splice forms A and B are the termed the flop and flip isoforms, respectively. Splice form C results in a frame shift, which probably results in a non-functional protein. Section B provides GenBank accession numbers for cDNAs or ESTs that demonstrate the existence of these three splice forms in mouse, rat, human, chicken and zebrafish. With the exception of GluR-C in the mouse and GluR-D in human, both flip and flop isoforms were observed in all species.

Figure 4.17. Locations of Known and Potential Editing Features on a Typical Ionotropic Glutamate Receptor.



The figure shows a typical ionotropic glutamate receptor, which includes NMDA, AMPA (GluR-A to D) and Kainate receptors (GluR-5&6). Each protein consists of four membrane regions (TM I-IV), the second of which enters and leaves the membrane on the same side. This results in an extra-cellular N-terminus and in intra-cellular C-terminus.

Each of the edited sites have been annotated as well as the flip/flop exon. The splicing of these two alternative exons may be regulated by editing of the flip branch site.

The diagram also indicates two additional regions where alternative splicing occurs, indicating that there are many different isoforms of this family of receptors.

protein was defined by the location of the ECD, which may contain altered amino acids.

Although the majority of known edited sites in the human genome are derived from inverted repeats[70], this is not a pattern we observed across our range of species. An ECD in a syntaxin exon (30th) was the only novel candidate exon in the top 50 that was derived from repeats. It has several A-G mismatches associated with it, suggesting that it may be functional. Repeats, especially simple repeats, are known to evolve relatively rapidly[178]. These observations suggest that widely conserved ECDs based on repeats may exist, but are likely to be rare.

To determine if there are any sub-categories of gene that are enriched in this set, we performed a test of statistical over-representation of gene ontology (GO) terms. This was performed using the online algorithm, FuncAssociate[179]. Unfortunately, the only GO term that showed any enrichment was the 'nucleus' GO term, which was observed in 17 of the genes, with a significance of 0.036 (once adjusted for multiple-testing). This is barely significant and not very interesting as it is such a broad term.

However, while investigating the individual functions of these candidates a few interesting observations were made. There were several candidate exons from genes involved in splicing, ranking 14th, 16th, 31st & 34/5th. This is particularly interesting given that there are now several lines of evidence that suggest editing and splicing interact (see Introduction). There were also several exons from genes involved in synaptic activity, ranking 22nd, 30th and 45th. These made convincing candidates given the observed bias towards editing in synaptic genes in both vertebrates and *Drosophila*[10,34,140]. The *Neuroligin 3* exon was of particular interest given that it is believed to interact with neurexins, and there are three relatively high-ranking neurexins in this data (ranking 100th, 101st & 106th).

Of the remaining exons, three were involved in neurogenesis (27th, 38th & 49th), two were involved in chromatin regulation (20th & 41st), and two were hnRNP proteins (9th & 36th), one of which is though to be involved in pre-mRNA splicing (hnRNP H). This last result was of particular interest as it has been shown that the ADAR enzymes interact with large nuclear RNP complexes[115]. Strictly speaking, the significance of each of these observations was unclear given that they were made without a statistical framework.

It was noticed that many of the genes with an exon in the top 50 had an additional high-ranking exon. Table 4.6 (117) describes all the genes that have more than one novel candidate exon in the top 100. Given that the known edited genes typically have more than one edited site, the nine genes listed in this table make very good candidates. The probability of two exons from the same gene in the top 36 novel candidates is approximately 0.05 (equation shown below). The significance of finding five such pairs of exons is much greater (roughly approximates to $3x10^{-7}$). Of course, such results do not necessarily predict RNA editing and may reflect other phenomena, such as alternative splicing.

$$\text{Probability of Finding Any Gene with 2 Exons in top 50} = P(1) \approx \frac{(\text{Av.Num.Exons per Gene}) \times (36-1)^2}{(\text{Total Number of Mouse Exons})}$$

$$\text{Probability of Finding 5 Genes With 2 Exons in top 50} \approx P(1)^5$$

4.6.5 Locations of the Predicted ECDs

Another observation that was made of the known edited sites was that they tend to overlap or occur near to the exon ends. Figure 4.9 (p96) shows that this is true for 9 of the 12 known edited sites. For this reason, the top 30 new candidates were sorted into appropriate categories to describe the locations of the two halves of their ECDs. The results of this are shown in Table 4.7 (p118). This table shows three categories of ECDs: standard, boundary and intronic.

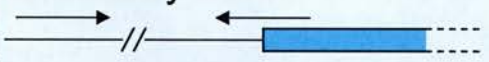
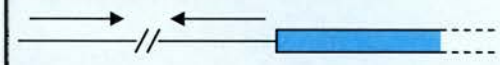
- ***Standard ECDs*** are those where the ES half of the ECD is entirely within the exon and the ECS half is entirely within the intron. This is the case for several of the known edited sites including the *GluR-B* Q/R site, the *BC10* site and the $5HT_{2C}R$ site.
- ***Boundary ECDs*** are those where the ES half of the ECD overlaps the exon boundary. The abundance of these sites, both in the candidates and in known edited exons, could be due to the reduced sequence constraints outside the exonic coding sequences, which could facilitate the evolution of ECD structures. Alternatively, these sites may also function in the regulation of the splicing of the transcript. Most of the known edited sites are in this category.

Table 4.6. Genes with ECD Predictions in Multiple Exons

Category	Ensembl Gene	Ensembl Exon	Primary Rank	Other Rank(s)	Ensembl Gene Description
Gene has two exons ranking 50th or better	ENSMUSG00000025255	ENSMUSE00000466378	17th	39th	Zinc finger homeodomain 4.
	ENSMUSG00000027018	ENSMUSE00000165486	20th	41st	Histone aminotransferase 1; histidine aminotransferase 1.
	ENSMUSG00000027108	ENSMUSE00000166139	37th	43rd	Putative GTP-binding protein PTD004 homolog.
	ENSMUSG00000028546	ENSMUSE00000375270	16th	34th	ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD) (Hu-antigen D).
	ENSMUSG00000033565	ENSMUSE00000463801	25th	33rd	RNA binding motif protein 9; fox-1 homolog (C. elegans); hexaribonucleotide binding protein 2; Fyn-binding molecule
Gene has one exon in top 50, & another exon in top 100.	ENSMUSG00000047022	ENSMUSE00000113629	10th	60th	Mirror-image polydactyly gene 1 protein homolog.
	ENSMUSG00000061603	ENSMUSE00000113867	13th	58th & 90th	NA
	ENSMUSG00000029563	ENSMUSE00000496440	15th	52nd & 107th	Forkhead box protein P2.
	ENSMUSG00000061455	ENSMUSE00000178321	30th	96th	Syntaxin variant

Genes with ECD predictions in multiple exons. Five genes have two exons in the top-scoring 50 exons. Four additional genes have one exon in the top 50 and another in the top-scoring 100 exons. The ranks of these exons are shown, with the primary rank being the rank of the best exon for that gene.

Table 4.7. ECD Locations for the Top 30 Predicted ECDs

ECS Type	Standard ECD	Boundary ECD	Intronic ECD
			
Known ECDs	Known ECDs in this category include 2 sites (5HT & BC10).	Known ECDs in this category include 7 sites (GluR-B Q/R, GluR-5 Q/R, GluR-6 Q/R, GluR-B R/G, GluR-C R/G, GluR-D R/G & ADAR2).	There are no known ECDs in this category.
Functions	Possible functions are probably restricted to recoding the protein encoded.	Possible functions could include both recoding of the protein encoded and regulation of splicing.	Possible functions are probably restricted to the regulation of splicing, although if the ECDs extend further than predicted, they may result in recoding of the encoded protein.

Top 30 Candidates

Rank	Exon ID	ES/ECS Separation	ES/ECS Locations	Rank	Exon ID	Exon Size	ES/ECS Separation	ES/ECS Locations	Rank	Exon ID	Exon Size	ES/ECS Separation	ES/ECS Locations
5th(K)	ENSMUSE00000223233	280bp	E/3'	1st(K)	ENSMUSE00000478281		4bp	3'/3'	2nd	ENSMUSE00000477286		0bp	5'/5'
14th	ENSMUSE00000186519	98bp	E/5'	4th(K)	ENSMUSE00000223221		17bp	3'/3'	3rd	ENSMUSE00000490396		0bp	5'/5'
15th	ENSMUSE00000496440	412bp	E/5'	6th(K)	ENSMUSE00000519849		11bp	3'/3'	7th	ENSMUSE00000105648		0bp	5'/5'
16th	ENSMUSE00000375270	2,309bp	E/3'	9th	ENSMUSE00000479443		0bp	5'/5'	8th	ENSMUSE00000515006		0bp	5'/5'
17/18th	ENSMUSE00000466378	501bp	E/3'	10th	ENSMUSE00000113629	34bp	16bp	3'/5'	22nd	ENSMUSE00000195189		0bp	3'/3'
19th(K)	ENSMUSE00000472446	450bp	E/5'	11th	ENSMUSE00000228059		0bp	3'/3'	23rd	ENSMUSE00000207263	94bp	1,915bp	3'/5'
20th	ENSMUSE00000165486	408bp	E/3'	12th	ENSMUSE00000177912	126bp	122bp	3'/5'	24th	ENSMUSE00000395043		0bp	3'/3'
28/29th	ENSMUSE00000469222	392bp	E/5'	13th	ENSMUSE00000113867		2,152bp	3'/3'	30th	ENSMUSE00000178321		0bp	3'/3'
				21st	ENSMUSE00000327165	135bp	145bp	3'/5'					
				25th	ENSMUSE00000463801	93bp	1,198bp	5'/3'					
				26th	ENSMUSE00000099319	39bp	53bp	3'/5'					
				27th	ENSMUSE00000153249	63bp	76bp	5'/3'					

Statistics

Average Rank	16.9					Average Rank	13.8					Average Rank	14.9
Median Distance	410bp					Median Distance	35bp					Median Distance	0bp
Total ECDs	8					Total ECDs	12					Total ECDs	8
Expected ECDs	14.6					Expected ECDs	6.7					Expected ECDs	6.7

$$Exp(\text{Standard}) = \sum_{i=0-28}\left(\frac{Size_i - 25}{Size_i + 75}\right) \qquad Exp(\text{Boundary}) = \sum_{i=0-28}\left(\frac{50}{Size_i + 75}\right) \qquad Exp(\text{Intronic}) = \sum_{i=0-28}\left(\frac{50}{Size_i + 75}\right)$$

Key

ENSMUSE00000469222 - AG Mismatch overlapping ES

ENSMUSE00000466378 - AG Mismatch within 15bp of ES

(K) - Exon is a known edited exon.

34bp 16bp - ES/ECS Separation and Exon Size are similar. The two halves of the ECD flank the exon closely.

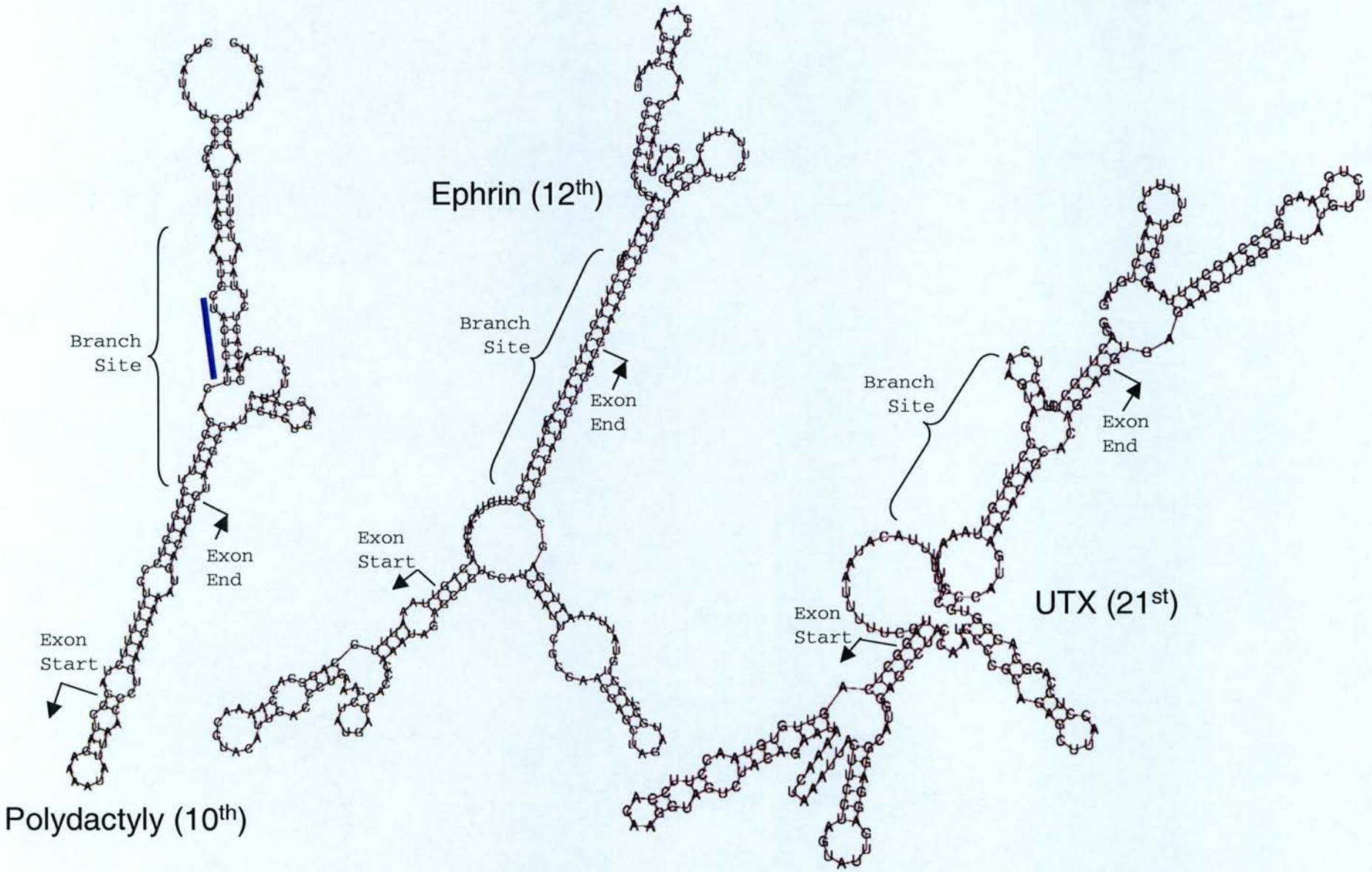
- ***Intronic ECDs*** were an unexpected product of these analyses. In order to detect boundary ECDs, the protocol used extended exons with an additional 50bp of flanking sequence at either end. Although this was successful, it had the side effect that some predicted ECDs were predicted where both ECD halves are located entirely in intronic sequences (i.e. the ES was in the flanking region, while the ECS was in one of the introns). Clearly these ECDs cannot result in recoding of amino acid sequences unless the ECD extends into the exon. It is also possible that these structures may be involved in the regulation of splicing. These ECDs are termed here as 'Intronic ECDs'. The four homologous predicted ECDs in the *GluR-A,B,C&D* receptor transcripts that are predicted to affect the 'flip' branch site are examples of this type of ECD.

Based on the length of each of the exons, it was possible to calculate the expected number of exons that should have occurred in each category. The results for this and the equations used are shown in Table 4.7 (p118). The boundary category had an excess over the expected, while the standard category showed a deficit. The significance of these values was tested using a standard chi-square test, resulting in a p-value of 0.17. This showed that, although the numbers indicated an excess of boundary ECDs, as seen with the known edited sites, this was not a significant result. Equally, there did not appear to be any significant difference between the average score ranks for each of these categories. Neither did there appear to be a significant bias towards 5 prime or 3 prime ECDs for each category.

4.6.6 Exons with Flanking ECDs

One interesting observation about this data was that there appeared to be five boundary ECD predictions that almost directly flanked their associated exons, with one ECS overlapping one exon end and the other ECS overlapping, or very near, to the other exon end. The function of this was not clear, however, it was tempting to think that the resulting duplexes may control splicing at both ends of the exon. In the hope of validating and elucidating the function of these structures, RNAFold predictions for the exons and neighbouring sequences were generated, as shown in Figure 4.18 (p120). These structures appeared to support the ECS predictions, with each exon showing significant secondary structure incorporating the exon ends. Possible splice branch sites were identified to examine the possibility that these

Figure 4.18. RNA Structure Predictions for Putative Flanking ECDs



structures may interfere with this step of splicing. In summary, these sites appeared very interesting, but it was not clear if they were involved in splicing, editing, both or some entirely different process.

4.6.7 Exonic ECSs

Due to the randomisation step in this protocol it was not possible to generate LOD scores for predicted ECDs that occur between two parts of the same exon. This is unfortunate as at least one known ECDs forms in this way (e.g. *KCNA1*). However, when reports were generated for each of the top 50 exons, ECSs formed in this way were occasionally more convincing than the ECSs formed using intronic sequence. These ECSs are shown in the figures in Appendix 1. These ECSs are termed as 'Exonic ECSs'.

4.6.8 Experimental Validation of Candidates

Many of the above candidates may be validated in the laboratory in collaboration with Mary O'Connell's RNA editing group. No results have been obtained at present, however, this process in ongoing.

4.6.9 A Smaller Range of Species

A drawback with this protocol is that it preferentially identified ECDs that are conserved over large evolutionary distances. This may miss ECDs that have only evolved relatively recently, such as in the mammalian lineage, or have been lost in more distant lineages. By ignoring the LOD scores between mouse and non-mammalian species, this analysis can be used to identify candidate mammalian specific ECDs.

Constraining the analysis to mouse, rat and human provided a method to identify mammal specific ECSs. The top 50 exons produced using this analysis are shown in Table 4.8 (p123). Many of the candidates obtained are the same as with the full vertebrate analysis, with 19 of the 50 exons being common to both top 50 lists. Ignoring the known edited genes, six of the remaining exons were from hnRNP genes, three were from synaptic genes, two were from chromatin regulation genes,

Table 4.8. ECS Predictions for the 50 Top-Scoring Exons in Mammals

Rank	Ensembl Exon	Ensembl Gene	Total LOD	Rat LOD	Hum LOD	Previous Rank	Putative Function	Ensembl Gene Description
1st	ENSMUSE00000472446	ENSMUSG00000057453	18.9	6.7	12.2	Pos 19th	-	Bladder cancer-associated protein (Bladder cancer 10 kDa protein) (Bc10).
2nd	ENSMUSE00000478281	ENSMUSG00000001986	15.9	5.0	10.9	Pos 1st	Synaptic	glutamate receptor, ionotropic, AMPA3 (alpha 3); cAMP-dependent Rap1 guanine-nucleotide exchange factor.
3rd	ENSMUSE00000110929	ENSMUSG00000014349	15.9	5.0	10.9	47th	-	NA - Involved in Ubiquitin conjugation
4th	ENSMUSE00000457409	ENSMUSG00000054423	15.5	5.8	9.7	59th	Neural?	Ca<2+>dependent activator protein for secretion; Ca2+-dependent activator protein for secretion.
5th	ENSMUSE00000440068	ENSMUSG00000053730	15.2	4.3	10.9	75th	-	NA - Transmembrane protein
6th	ENSMUSE00000477286	ENSMUSG00000033981	14.7	5.0	9.7	ABCD 2nd	Synaptic	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
7th	ENSMUSE00000397418	ENSMUSG00000017740	14.7	5.0	9.7	83rd	Synaptic	Solute carrier family 12 member 5 (Electroneutral potassium-chloride cotransporter 2) (K-Cl cotransporter 2)
8th	ENSMUSE00000490396	ENSMUSG00000001986	14.7	5.0	9.7	ABCD 3rd	Synaptic	glutamate receptor, ionotropic, AMPA3 (alpha 3); cAMP-dependent Rap1 guanine-nucleotide exchange factor.
9th	ENSMUSE00000133193	ENSMUSG00000056851	14.7	5.0	9.7	82nd	hnRNP	Poly(rC)-binding protein 2 (Alpha-CP2) (Putative heterogeneous nuclear ribonucleoprotein X) (hnRNP X)
10th	ENSMUSE00000223233	ENSMUSG00000033981	14.7	5.0	9.7	Pos 5th	Synaptic	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
11th	ENSMUSE00000234858	ENSMUSG00000036208	14.1	1.9	12.2	99th	-	NA
12th	ENSMUSE00000223221	ENSMUSG00000033981	14.0	4.3	9.7	Pos 4th	Synaptic	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
13th	ENSMUSE00000382622	ENSMUSG00000021111	13.6	5.0	8.5	111th	Polymerase A	Poly(A) polymerase alpha (EC 2.7.7.19) (PAP) (Polynucleotide adenylyltransferase).
14th	ENSMUSE00000188013	ENSMUSG00000000568	13.6	5.0	8.5	110th	hnRNP	Heterogeneous nuclear ribonucleoprotein D0 (hnRNP D0) (AU-rich element RNA-binding protein 1).
15th	ENSMUSE00000435727	ENSMUSG00000033981	13.3	3.6	9.7	Pos 80th	Synaptic	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
16th	ENSMUSE00000153205	ENSMUSG00000025788	13.3	3.6	9.7	121st	Chromatin	NA - Putative chromatin helicase
17th	ENSMUSE00000335487	ENSMUSG00000033981	13.3	3.6	9.7	Pos 122nd	Synaptic	Glutamate receptor 2 precursor (GluR-2) (GluR-B) (GluR-K2) (Glutamate receptor ionotropic, AMPA 2).
18th	ENSMUSE00000373990	ENSMUSG00000021912	13.2	6.7	6.5	124th	Chromatin	Polybromo 1. Involved in chromatin remodelling.
19th	ENSMUSE00000289393	ENSMUSG00000044465	12.8	4.3	8.5	135th	-	NA
20th	ENSMUSE00000314161	ENSMUSG00000021028	12.8	4.3	8.5	87th	-	MAP3K12 binding inhibitory protein 1 (MAPK upstream kinase-binding inhibitory protein)
21st	ENSMUSE00000103873	ENSMUSG00000007850	12.8	4.3	8.5	136th	hnRNP	Heterogeneous nuclear ribonucleoprotein H (hnRNP H).
22nd	ENSMUSE00000426424	ENSMUSG00000028546	12.8	4.3	8.5	34th	mRNA splicing	ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD) (Hu-antigen D).
23rd	ENSMUSE00000202102	ENSMUSG00000030697	12.8	4.3	8.5	138th	Microtubule re-organisation?	Serine/threonine protein phosphatase 4 catalytic subunit (EC 3.1.3.16) (PP4C) (Pp4) (Protein phosphatase X)

Table 4.8. ECS Predictions for the 50 Top-Scoring Exons in Mammals

Rank	Ensembl Exon	Ensembl Gene	Total LOD	Rat LOD	Hum LOD	Previous Rank	Putative Function	Ensembl Gene Description
24th	ENSMUSE00000133722	ENSMUSG00000023150	12.8	4.3	8.5	81st	RNA splicing	influenza virus NS1A binding protein.
25th	ENSMUSE00000342145	ENSMUSG00000028546	12.8	4.3	8.5	35th	mRNA splicing	ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD) (Hu-antigen D).
26th	ENSMUSE00000457469	ENSMUSG00000037369	12.8	4.3	8.5	137th	-	Ubiquitously transcribed X chromosome tetratricopeptide repeat protein (Ubiq. transcribed TPR protein on the X
27th	ENSMUSE00000375270	ENSMUSG00000028546	12.8	4.3	8.5	16th	mRNA splicing	ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD) (Hu-antigen D).
28th	ENSMUSE00000155429	ENSMUSG00000060679	12.8	4.3	8.5	36th	hnRNP	mitochondrial ribosomal protein S9; muscle protein 637.
29th	ENSMUSE00000287940	ENSMUSG00000031302	12.8	4.3	8.5	45th	Synaptic	Neuroligin 3 precursor (Gliotactin homolog).
30th	ENSMUSE00000462487	ENSMUSG00000021546	12.8	4.3	8.5	133rd	hnRNP	Heterogeneous nuclear ribonucleoprotein K.
31st	ENSMUSE00000472905	ENSMUSG00000031012	12.8	4.3	8.5	117th	Possibly Synaptic	Peripheral plasma membrane protein CASK (EC 2.7.1.-) (Calcium/calmodulin-dependent serine protein kinase).
32nd	ENSMUSE00000460496	ENSMUSG00000061731	12.8	4.3	8.5	134th	-	Exostosin-1 (EC 2.4.1.224) (EC 2.4.1.225)
33rd	ENSMUSE00000479443	ENSMUSG00000007850	12.8	4.3	8.5	9th	hnRNP	Heterogeneous nuclear ribonucleoprotein H (hnRNP H).
34th	ENSMUSE00000504265	ENSMUSG00000020385	12.6	4.6	8.0	31st	mRNA splicing	Dual specificity protein kinase CLK4 (EC 2.7.1.37) (EC 2.7.1.112) (CDC like kinase 4).
35th	ENSMUSE00000147930	ENSMUSG00000025092	12.5	0.4	12.2	144th	-	Heat shock 70 kDa protein 12A.
36th	ENSMUSE00000485910	ENSMUSG00000025782	12.5	5.0	7.5	149th	-	TAFII140 protein (Fragment). TATA binding protein
37th	ENSMUSE00000457290	ENSMUSG00000053716	12.5	5.0	7.5	148th	-	Dual specificity protein phosphatase 7 (EC 3.1.3.48) (EC 3.1.3.16).
38th	ENSMUSE00000487767	ENSMUSG00000052534	12.5	5.0	7.5	147th	-	Pre-B-cell leukemia transcription factor-1 (Homeobox protein PBX1).
39th	ENSMUSE00000318616	ENSMUSG00000052834	12.3	4.3	8.0	160th	-	NA
40th	ENSMUSE00000163770	ENSMUSG00000026874	12.2	NA	12.2	165th	-	Complement C5 precursor (Hemolytic complement) [Contains: C5a anaphylatoxin].
41st	ENSMUSE00000511130	ENSMUSG00000054640	12.2	NA	12.2	164th	-	Sodium/calcium exchanger 1 precursor (Na(+)/Ca(2+)-exchange protein 1).
42nd	ENSMUSE00000452455	ENSMUSG00000054640	12.2	NA	12.2	163rd	-	Sodium/calcium exchanger 1 precursor (Na(+)/Ca(2+)-exchange protein 1).
43rd	ENSMUSE00000211495	ENSMUSG00000031654	12.1	3.6	8.5	167th	Synaptic	Cerebellin precursor (Precerebellin) (Brain protein D3).
44th	ENSMUSE00000106115	ENSMUSG00000018651	12.1	3.6	8.5	166th	-	transcriptional adaptor 2 (ADA2 homolog, yeast)-like.
45th	ENSMUSE00000306694	ENSMUSG00000025278	12.0	5.0	7.0	56th	-	Filamin B (FLN-B) (Beta-filamin) (Actin-binding like protein) (ABP- 280-like protein).
46th	ENSMUSE00000410358	ENSMUSG00000010797	12.0	5.0	7.0	175th	-	Wnt-2 protein precursor (IRP protein) (INT-1 related protein).

Table 4.8. ECS Predictions for the 50 Top-Scoring Exons in Mammals

Rank	Ensembl Exon	Ensembl Gene	Total LOD	Rat LOD	Hum LOD	Previous Rank	Putative Function	Ensembl Gene Description
47th	ENSMUSE00000379978	ENSMUSG00000052063	11.9	3.9	8.0	179th	-	NA
48th	ENSMUSE00000340695	ENSMUSG00000022708	11.8	4.3	7.5	189th	-	zinc finger protein 288; POZ/zinc finger transcription factor; POZ/zinc finger transcription factor ODA-8.
49th	ENSMUSE00000100364	ENSMUSG00000020074	11.8	4.3	7.5	95th	-	cell division cycle and apoptosis regulator 1.
50th	ENSMUSE00000224458	ENSMUSG00000027168	11.8	4.3	7.5	44th	-	Paired box protein Pax-6 (Oculorhombin).

The 50 top-scoring exons in the screen of three mammalian species. Each exon is listed with its Ensembl exon and gene identifiers and the gene description. A high combined LOD score indicates a convincing conserved ECD structure. Individual LOD scores are provided for both the mouse:rat and the mouse:human comparison. Where the exon is a positive control the row has been shaded a darker blue. The two exons that have previously been suggested to contain a duplex that covers the flip branch site have also been shaded dark blue and annotated as 'ABCD' in the 'Previous Rank' column. Where the exon was found in the top 50 for the full vertebrate screen, the row has been shaded purple. Novel candidate exons are shaded light blue. Where the exons are from genes that are known to contain recoding edited sites, they are annotated as 'Pos' in the 'Previous Rank' column. The putative function column contains brief information on the putative functions of these genes. Primarily these include: Splicing machinery, Synaptic machinery, hnRNPs and Chromatin regulation.

and five were involved in splicing. Although some of these genes were common to both analyses, many of them are not. Once again, it was unclear if this apparent enrichment was significant. Another interesting observation from these mammalian ECD predictions is that there is a different Polymerase Alpha gene candidate instead of the one found in the full vertebrate analysis. Unfortunately, such speculation can only be confirmed by laborious wet-lab techniques, and as such, many of these possibilities will not be investigated further.

4.7 Conclusion

This chapter has described the first protocol for identifying conserved edited sites through their ECDs. An outline of this method is given in Figure 4.2 (p80). This method has been used to specifically identify ECDs for the known edited site, to screen a range of six vertebrate species for these and novel ECDs and finally, to screen a range of three mammalian species. The protocol has been successful in each of these cases. Of the ten known protein recoding edited sites with experimentally validated ECD structures, this protocol was able to correctly identify nine of these ECDs as the best ECD for each given exon. Each of these ECDs was found in three or more species, and two of them were found in all six species (*GluR-B&C* R/G sites). This represents an evolutionary distance of 450million years of divergence[86]. The *KCNA1* ECD was the only one not identified by this method. Based on this success, predictions were made for the known sites that did not have experimentally validated ECD structures. Two of these sites had published MFOLD RNA structure predictions, which were confirmed by this protocol. The remaining known edited sites also had ECDs predicted, although, with the exception of the *GluR-6* I/V ECD, they were not as convincing as those seen for the published ECDs. A more sensitive approach did not identify any additional ECDs.

The method was then used to predict ECDs conserved between every mouse exon and their orthologous exons in the five other species (rat, human, chicken, zebrafish and pufferfish). The results were then ranked by the quality of their conserved ECDs. Nine of the known edited sites were found to rank in the top 0.5% of all mouse exons, including the ranks of 1st, 4th, 5th, 6th, 19th and 42nd. This demonstrated that this method was sensitive and specific enough to use on whole genomes. Four of the novel ECD predictions, which turned out to be homologous, were found in *GluR-A,B,C&D* exons. These ECDs show remarkable sequence and structure conservation both between the four genes and between the multiple species in which they were observed. This structure, which was conserved in all species for at least one of the homologues, appears to cover the branch site of the flip exon of these genes. Alternative splicing of these genes is known to have a major effect on the properties of the channel produced by their proteins. It was tempting to think that this conserved ECD may result in editing of the branch point and force the other exon, flop, to be spliced instead of the flip exon. Five of the other conserved ECDs appear to flank their exons, suggesting that they may regulate splicing at both ends of their exons. It

was also noted that many genes had more than one exon in either the top 100 or the top 50, which makes these candidates considerably more convincing.

This method was also applied to predict mammalian specific ECDs. The results from this screen were broadly similar to the screen across all six vertebrates, although some results differed. Both of these screens identified a number of ECDs in genes involved in splicing, synaptic activity and hnRNPs.

Without thorough experimental validation it is not clear how many novel edited sites can be confirmed using this protocol. However, this screen can be compared to previous screens based on its ability to identify the known edited sites. The only two screens that have successfully identified any of the known recoding edited sites are the screen in the previous chapter[79] and the screen by Levanon *et al*[81]. The latter screen identified several of the edited sites in glutamate receptors, although it did not state which ones, as well as identifying *Flna, BC10* and *Cyfip2*. The mismatch screen identified the *BC10* site, the *GluR-B* Q/R site, the *GluR-C&D* R/G sites and the *5HT$_{2C}$R* site in the top 6 genes (see Chapter 3). This screen has identified the three *GluR-B,C&D* R/G sites, the *GluR-B* Q/R site, the *BC10* site and the *GluR-6* Q/R site in the top 42 exons (29 genes). Each of these screens has identified 5 or 6 of the known edited sites, with only a small number of additional predictions at the given cut-offs. This shows that these screens are comparable. Interestingly, the specific sites that are identified by each screen overlap considerably, however, the additional predictions do not appear to overlap.

4.7.1 Caveats and Restrictions of the Protocol

There are a number of reasons why this protocol would not find some edited sites. The nature of this protocol means that ECDs can only be identified in exons that are annotated in Ensembl and have orthologous genes predicted in at least one of the other species. Any ECDs that occur entirely in the middle of introns or in exons that are not annotated would be missed. Due to the randomisation procedure, ECDs that form entirely within exons are generally missed as well.

It is also possible that there are some sites that are not mediated by a normal ECD structure, and that the duplex is formed through interaction with different molecules (e.g. anti-sense transcripts). Even if we assume that edited sites contain ECDs, these

structures may not be conserved, either because editing is not conserved, or an alternative ECD is used in other species. Other ECDs may be poorly conserved, due to poor conservation of sequence, secondary structure or both. The ECD may also be too complex to identify using this protocol. For example the *synaptotagmin-I* ECD in *Drosophila* has been shown form a long-range pseudo-knot[108], which would probably not be detected by this method. Alternatively, the ECD may be shorter than the scan size of 25bp, the ECS may be more than 2.5kb from the exon, or the ECS may be in a more distant intron. Finally, editing could occur in the mature transcript in some cases, given that ADAR1 shuttles into the cytoplasm[180,181]. Future work could be aimed at investigating these possibilities.

Despite these issues, however, this protocol appears to have been successful both at identifying novel candidate ECDs for known sites and for identifying novel edited sites through their ECDs.

5 Results: Conserved RNA Duplexes in The Fruit Fly

5.1 Preface

The methods used in the previous chapter are theoretically applicable to any group of two or more species for which genomes and gene annotation are available. As long as ECD structures are sufficiently conserved between the two species, this protocol should be successful. The fruit fly (*Drosophila*) has been shown to be a good model organism for studying A-I RNA editing. There are many known edited sites, most of which are involved in nervous system function and integrity[34]. *Drosophila* species contain only one ADAR enzyme, *dADAR*, in comparison to mammals, which have three (*ADAR1, ADAR2 & ADAR3*)[11]. This makes it simpler to probe exactly how A-I editing operates. Flies are also relatively inexpensive, require minimal paperwork and are easy to genetically manipulate. Additionally, Mary O'Connell's group has worked on editing in the fruit fly for over five years (MRC Human Genetics Unit). For these reasons, a screen for ECDs in *Drosophila* was considered an important analysis.

5.2 Introduction

Initially, a comparison of mouse and *Drosophila melanogaster* was performed using the same protocol as described in the previous chapter. Unfortunately, it appeared that the evolutionary distance between these two species was too great to be able to identify conserved ECDs. This screen did not identify any of the known edited exons from either species. In agreement with this observation, only one of the known mammalian edited sites, the *KCNA1* I/V site, can also be observed in the fruit fly[80,135,140]. In this case, the same amino acid change is observed in the homologous *Drosophila Shab* gene as the mammalian *KCNA1* gene. The *Drosophila Shaker* gene, which is orthologous to *KCNA1*, is also edited, but in a different position[80].

However, Hoopengardner *et al* have shown that a comparison of two *Drosophila* species can identify edited sites with high sensitivity and specificity[140]. These two species, *Drosophila melanogaster* (*D.mel*) and *Drosophila pseudoobscura* (*D.psu*), both have near-complete genome sequences and gene annotation, which means that

they are suitable for this type of analysis. A further consideration was that the number of ESTs or cDNAs available for *Drosophila* species is limited. This means that it would be difficult to identify edited sites using a method based on mismatches, although one edited site, *Ca-alpha-1T*, was initially identified in this way[135]. Although there are many known *Drosophila* edited sites, very few of them actually have known ECD structures (see Table 5.1 – p132). There are sixty-five known locations of individual or clusters of edited sites. However, only seven of these have published ECD predictions[108,113,139,182]. Only two of these have been experimentally validated[108]. In addition to identifying potentially novel editing sites, this protocol could identify ECD structures for the known edited sites. As a result of these observations, this chapter concentrates on the identification of conserved ECD predictions based on these two species.

Table 5.1. The Known *Drosophila* Edited Sites

Ensembl Gene	Reference	Edited Site	Ensembl Exon	Published ECD	Edited Sequence (A = Edited Adenosine)
CG1522	Smith *et al*[183]	Cac K/K (12)	CG1522:11	-	aacgcugccAAA
		Cac 15	CG1522:13	-	gcguugcauuA
		Cac 17a	CG1522:15	-	guauucacgauAg
		Cac 17b	CG1522:15	-	agagaauucuggaAua
		Cac 17c	CG1522:15	-	aauguuguuaAcau
		Cac 17d	CG1522:15	-	guacggacgaaA
		Cac 19	CG1522:17	-	uuccggaucgaaA
		Cac 24	CG1522:21	-	ucgaAcucaaucaac
		Cac 26a&b	CG1522:23	-	ggaauaugauccaAA
		Cac 30b	CG1522:27	-	cgcgguauccaAga
		Cac 30a	CG1522:27	-	uaccguaucuAau
CG9907	Hanrahan *et al*[184]	Para t/m	NULL:512413	Predicted[113]	cAauuaauugaAAau
		Para ssp	CG9907:26	Predicted[139]	uuuuugcuggaaAaua
		Para sfc	CG9907:28	Predicted[139]	cuauaAugcuaug
		Para fsp	NULL:512549	Predicted[139]	ccaaugcgcAggcuc
CG7535	Semenov *et al*[185]	GluCl alpha-1 78/79	CG7535:3	-	uaaugccaaAA
		GluCl alpha-1 722	CG7535:8	-	gaucuacuauucaA
		GluCl alpha-1 1034	CG7535:10	-	ugcauaaggagaA
		GluCl alpha-1 1179	CG7535:11	-	aaaagcggcuccaAug
CG4128	Grauso *et al*[186]	Dalpha6 Cluster	CG4128:6	-	acguaucacaccaAcAuuguggucaaacAuAAcggca
CG12598	Palladino *et al*[5]	dADAR	CG12598:9	Prediction[182]	acggauauuuAgucc
CG18314	Stapleton *et al*[135] & Xia *et al*[133]	CG18314 Clust 3	CG18314:5	-	uaAgacuaAgggccguuugcaagcagagcuAAucg
		CG18314 Clust 4	CG18314:5	-	cagaugacuAgaguua
		CG18314 Clust 1	CG18314:5	-	aaaAuuuAucgaacccuAgucauau
		CG18314 Clust 2	CG18314:5	-	gucgucuggAuc
CG15899	Hoopengardner *et al*[140]	CaAlpha1T	CG15899:15	-	uugagggauucAgu
CG4894	Hoopengardner *et al*[140]	DmCa1D Cluster	CG4894:5	-	uuuAgccauaugguuuuguguuacauAAuggugcauaucuAA
		DmCa1D Last Site	CG4894:5		auuagauuuuacaauuguaguuauAgg
CG12295	Hoopengardner *et al*[140]	Alpha2Delta I/Va	CG12295:5	-	gacagaggcuagauAu
		Alpha2Delta I/Vb	CG12295:5	-	gacuuugugaacAu
		Alpha2Delta R/G	CG12295:6	-	auugaacuguuuAg
CG12348	Hoopengardner *et al*[140]	Shaker KERG	NULL:538653	-	uuuagugaagaaauaAAa
		Shaker I/M	NULL:538657	-	uccuuggcaauAuu
		Shaker I/V	CG12348:15	-	uuugugcgugAucgc
		Shaker TA & QR	CG12348:15	-	gaaAcggaucA
CG10952	Hoopengardner *et al*[140]	EAG K/Ra	CG10952:11	-	gacaacgagaAgg
		EAG Y/C	CG10952:13	-	guacuaaacuAu
		EAG N/D & Silent	CG10952:13	-	guAuuuAacgagcauccg
		EAG AA	CG10952:13	-	uccggcAuuucgucu
		EAG K/Rb	CG10952:14	-	uuuuucgcaAgg
CG10693	Hoopengardner *et al*[140]	Slowpoke N/D	CG10693:6	-	gcuggauuuuAauaau
		Slowpoke S/G	NULL:1763569	-	uggucaaugauAgu
CG3139	Hoopengardner *et al*[140]	Syt I/V - A	NULL:915362	-	auaugugaaaAuu
		Syt K/R & I/V - B/C	NULL:915362	Validated[108]	aagaAgaagacaaguAucaaaaaaug
		Syt I/M - D	NULL:915362	Validated[108]	uugaacaaauAcaa
CG2999	Hoopengardner *et al*[140]	Unc-13 S/G	NULL:2361787	-	gauguuguuAgc
CG40306	Hoopengardner *et al*[140]	StnB T/A	CG40306:8-B	-	caggucucccauAcc
CG32490	Hoopengardner *et al*[140]	Cpx I/M NDSG	CG32490:6	-	aucaaauAgaaacgcaaguaAAugagcuaaaa
CG2520	Hoopengardner *et al*[140]	Lap T/A	CG2520:10	-	cuagcgcugcacuaAc
CG32975	Hoopengardner *et al*[140]	Da5 I/V (Nic34E)	CG32975:5	-	auuuuAAuauu
		Da5 T/A I/V L/L (Nic34E)	CG32975:6	-	uucAcaAuauuAgccacaauAgcuguacuacuAu
CG11348	Hoopengardner *et al*[140]	ARD R/G 64b	CG11348:3	-	guuggaguAAgauuu
		ARD I/M 64b	CG11348:4	-	aaucaAAuuaugaaa
CG6798	Hoopengardner *et al*[140]	SBD T/A & Silent (96A)	CG6798:7	-	guuAcguuAugu
CG10537	Hoopengardner *et al*[140]	Rdl L/L R/G	CG10537:4	-	cucguuuAgcguauAga
		Rdl I/V	CG10537:7	-	cugaucguuAuuauaucau
		Rdl N/D	CG10537:7	-	aucgcAaugcaacgccgg
		Rdl M/V	CG10537:8	-	cgaauucaAAugcgaaaac
CG1066	Bhalla *et al*[80]	Shab I/V	NULL:368787	-	guAAucgcuuug
CG13167	Xia *et al*[133]	CG13167 I/V	CG13167:1	-	cgauauccugcccAuc
CG9619	Xia *et al*[133]	CG9619 S/G	CG9619:3	-	gucggauugcuaAg
CG8428	Xia *et al*[133]	CG8428 N/G Spin	NULL:2096895	-	gcuuuaaauuugAAu
CG12076	Xia *et al*[133]	CG12076 Q/R	CG12076:3	-	gguuagcugcugcccA
CG14936	Xia *et al*[133]	CG14936 V/V	CG14936:3	-	cgugcgcuguguAa
		CG14936 G/G	NULL:1219665	-	uguggAcucacggacgau

5.3 The Known *Drosophila* Edited Sites

The first edited sites in the fruit fly were identified through serendipitous observations of A-G mismatches in cDNAs. This includes sites in *para* (paralytic)[184], *cac* (cacophony)[183], *Dα6*[186] and *GluCl-α1*[185]. A self-editing site in the only *Drosophila* Adar gene was also identified[5].

Since then there have been several directed attempts to identify additional edited targets in the *Drosophila* transcriptome. One of these methods, which relied on the identification of cDNAs containing A-G mismatches, resulted in the initial identification of the *Ca-alpha-1T* edited site[135]. They also identified a number of other putative edited candidates, however these sites have not been verified.

A more recent paper by Hoopengardner *et al*, used a very different, but highly effective approach to identifying novel edited sites in the fly. They scanned a collection of ion channels, G-protein coupled receptors, synaptic proteins and transcription factors for regions of exceptional sequence conservation between *D.mel* and *D.psu*[140]. Table 5.1 (p132) demonstrates the success of this approach as most of the currently known sites were identified in this screen, including several that were known before this screen. An additional editing site in the *Shab* gene was then identified via its homology to one of these sites (in the *Shaker* gene)[80]. Clearly the use of sequence conservation is a powerful method for identifying edited sites between these species.

Finally, Xia *et al* identified five novel edited sites using a hybrid approach based on the use of a inosine antibody and mismatch data[133].

There are several observations that indicate differences between the *Drosophila* and vertebrate edited sites. There are more known *Drosophila* genes with edited sites, and, on average, each gene contains many more edited sites. Only a small proportion of the *Drosophila* sites have published ECD predictions, compared to the majority of the vertebrate sites. Although the genes typically have neurological functions, the exact processes involved differ between *Drosophila* and vertebrates. For example, several *Drosophila* synaptic genes are known to be edited, but no vertebrate ones are known.

The strict approach that Hoopengardner *et al* used suggests that there could be many more *Drosophila* edited sites that remain unidentified. We aimed to identify some of these unknown sites by applying our ECD protocol to these species.

5.3.1 Modifications to the ECD Finding Protocol

There were several modifications to the previous protocol. In Chapter 4 LOD scores were required to combine the ECD scores from each species pair. In order to do this, randomisations were performed to generate a negative ECD score distribution. These randomised negative controls were not suitable for exonic ECDs, however. In the *Drosophila* protocol we only used two species, which meant there was no requirement for LOD scores or randomised controls. This in turn meant that ECDs where the ECS is located in the exon could be included in the analysis. The high levels of background conservation often seen in exons meant that these ECD predictions had to be ranked separately from the intronic ECD predictions.

Given the success of the Hoopengardner approach, it was decided to increase the importance of sequence conservation in this analysis. This was achieved by altering the scoring system, such that putative secondary structure was less important than sequence conservation (see Materials & Methods section). The ECD core size was also varied in an attempt to tailor these analyses to the fruit fly.

5.4 Full Protocol Description

The majority of this protocol is similar to that of the vertebrate screen. Where changes were made they are described here.

5.4.1 Data Preparation

5.4.1.1 Initial Files

A number of files from external sources were required for these analyses. These are detailed in this section.

- *Drosophila melanogaster* full genome sequence was obtained from Ensembl (version DROM 3A).
- *Drosophila pseudobscura* full genome sequence was obtained from FlyBase (version R1.03).
- *Drosophila melanogaster* exon sequences and coordinates were obtained from Ensembl Mart for all predicted genes (November 2004). (See Materials & Methods).
- Preliminary *Drosophila pseudobscura* gene and exon annotation was obtained from FlyBase for all genes (21/10/2004, Chado R1.03). Putative exon sequences were obtained from the genomic sequences based on the coordinates derived from this file. Exons were only included if they were contained within a predicted gene (as defined in the annotation file), even if the coordinates did not precisely agree. Orphan exons were excluded.
- *Drosophila melanogaster* gene descriptions were obtained from Ensembl Mart (November 2004). (See Materials & Methods).

The genomic sequence files were formatted into BLAST databases using the formatdb program from the BLAST package (version 2.2.6). Separate sequence indexes were generated for both the genomic and the exon FASTA files (as required for Sgrab – see Section 2.5.1). This allowed for rapid sequence retrieval using the Sgrab perl program.

5.4.1.2 Orthologous Gene Predictions

In contrast to the previous protocol, pre-computed orthologous gene predictions were not available between these two species. Instead, a reciprocal BLAST method was used to identify putative orthologous gene pairs. Similar protocols underlie orthology prediction in Ensembl. Concatenated exon sequences were generated for each predicted gene from both species. The sequences were then formatted into two BLAST databases, one for each species. The *D.mel* concatenated sequences were then BLAST searched against the *D.psu* database. The *D.psu* concatenated sequences were then BLAST searched against the *D.mel* database. The BLAST program used was MegaBLAST with the following options (-p blastn –W 11 –n T –m 8). This forced MegaBLAST to use a word-size of 11, which gives it the same sensitivity as a normal nucleotide-nucleotide BLAST search. BLAST matches with an E-value greater than 10^{-15} were ignored. Reciprocal BLAST hits are those where the best hit against one database, finds the target to be it's best hit in the other database (i.e. their respective best hit is each other). Gene pairs fitting these criteria were taken to be orthologous.

5.4.1.3 Orthologous Exon Predictions

The methods for obtaining orthologous exon predictions are identical to those in the previous chapter (Section 4.4.1.2). The BLAST match nucleotide identify percentage had to be above 60% for exons to be considered orthologous.

Unfortunately, many of the *D.mel* exons did not have orthologous exons predicted after this analysis (17%). This was primarily due to incomplete annotation of the *D.psu* genes and exons. In order to identify orthologous exons for these exons, a second screen based on a simple nucleotide BLAST search was carried out. All *D.mel* exons that did not already have an orthologous exon assigned were BLAST searched against the *D.psu* genome. The options used were (-m 8 –e 10 –r 1 –q –1 –G 2 –E 1 –W 11). These options modified the match, mismatch and gap penalties, specified a word length of 11, and restricted the output to tabular results with E-values less than 10. Modifying the gap penalties had the result that, where possible, the entire exons aligned, instead of the best conserved parts. Reciprocal BLAST hits were identified (see Section 2.4.2) for each *D.mel* exon. As long as the reciprocal

BLAST matches had bit scores greater than 50, they were considered to represent a putatively orthologous exon pair. Each putative orthologous *D.psu* exon was then given a unique identifier of the form GA999999--, so that they were easily distinguished from orthologous exons identified using the reciprocal BLAST method. These criteria were purposefully non-conservative, with the intention of maximising the number of exons for which orthologues are available. In addition to the 43,400 orthologous exon predictions previously obtained, this method produced just over 8,900 orthologous exon predictions.

5.4.2 Changes to the Main Program

With the exception of steps 8 and 9 (which deal with the LOD scores), the method is very similar to that described in Figure 4.2 (p80). In an effort to reduce the number of false predictions, the size of intron that was searched was reduced to 2kb. This was based on some of the locations of some of the published *Drosophila* ECD predictions, although this did not include the *syt* ECDs as they were published after this work was commenced. The importance of the sequence conservation in the predicted ECDs was increased by multiplying the appropriate scores in the ECD scoring system by the conservation bias. These included all the scores under the 'Exon conservation' and 'Intron conservation' sub-heading in Section 4.2.4.4. After experimentation with the known *Drosophila* edited sites, the value of the conservation bias was empirically set to 4 (unless otherwise stated). Once each score had been calculated, the putative core ECD alignments were recorded for future observation and analysis. Figure 4.2 (p80) provides an example of an ECD alignment with the core section highlighted in red (between steps 6 & 7).

5.4.2.1 Program Variables

In addition to the program variables described in Section 4.2.4.5, three more input variables were added. These are listed below.

Restrict ECS Location

The program can be restricted to only analyse ECSs located in the 5 prime, 3 prime, exonic or any combination of these three locations.

Sequence Conservation Bias

This variable increases the penalties against sequence mismatches in ECDs. This does not affect the scoring of the secondary structure of the ECDs.

Output Threshold Score

Only ECDs scoring below a desired threshold are included in the output.

5.4.3 Annotation of the Putative Conserved ECDs

In the previous chapter, it was necessary to generate a negative control data set and apply LOD scores to the data. This was not required in this protocol as there were only two species being analysed. As such the ECD scores alone are sufficient to rank the results. However, a number of further analyses were carried out to further characterise the ECDs of interest, which essentially consists of 3 groups of exons;

1. The known edited exons.
2. The top scoring candidate ECD predictions identified with ECSs in 5 prime or 3 prime intronic sequences.
3. The top-scoring candidate ECD predictions identified entirely within exonic sequences.

5.4.3.1 Identifying A-G Mismatches

To support the ECD predictions, all the Ensembl *D.mel* exons were BLAST searched against all the publicly available transcribed *Drosophila* sequences. These sequences were obtained from GenBank Entrez using the query "*Drosophila melanogaster* [ORGN] AND (cDNA [TITL] OR mRNA [TITL] OR gbdiv_est [PROP])" and limited to "mRNA". This resulted in ~420,000 expressed sequences. These exons were also BLAST searched against the *D.mel* genome. BLAST hits with an E-value below 10^{-9} were scanned for A-G mismatches. This threshold was empirically derived by examining alignments by eye. Any A-G mismatches that were observed in the expressed sequences, but not in the genomic sequences were recorded as potential editing events.

Finally, coding sequences were obtained (as described in Section 4.2.5.2) and alignment reports were generated (as described in Section 4.2.5.3.

5.5 Results for Known Edited Sites

5.5.1 Protocol Calibration

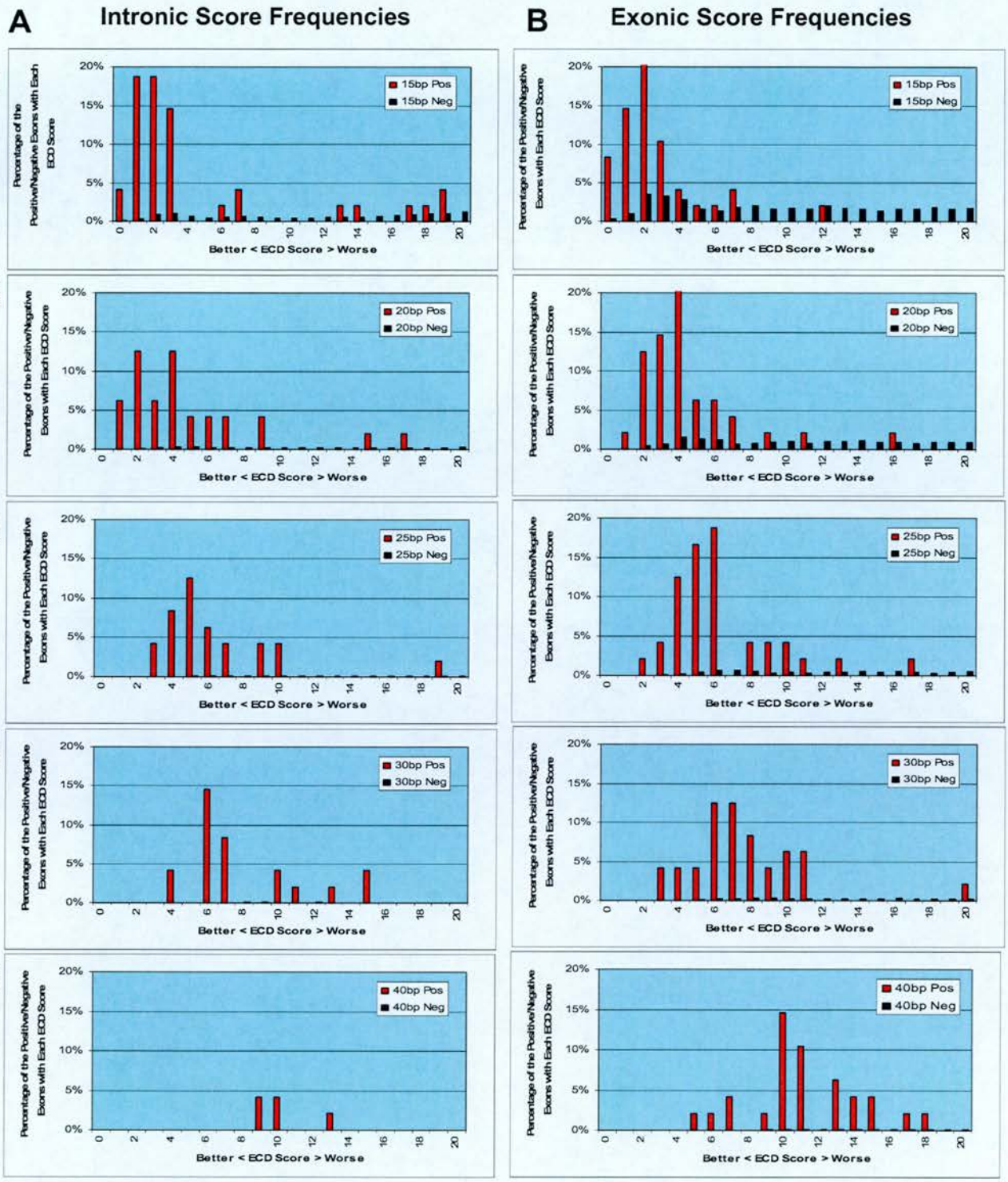
The known edited sites, as described in Table 5.1 (p132), were used to calibrate the variables available to this protocol, including sequence conservation bias and core ECD size. In order to rate the performance of each variable combination, we made the assumption that all exons that are not known to be edited, are not edited. This provides us with defined sets of positive and negative controls, from which estimates of specificity and sensitivity were calculated. Clearly there were likely to be additional edited sites, so these measures may give a conservative view of the performance of the protocol. However, given that we did not know what the additional sites were, it was not possible to make these estimates in any other way. As long as edited exons are relatively rare, these estimates of specificity and sensitivity should be roughly correct.

Initially, the protocol was used to identify ECDs with a core length of 25bp, which is the size used to identify mammalian ECDs in the previous chapter. Based on the specificity of the protocol, it was clear that raising the sequence conservation bias significantly improved the results of protocol. This observation makes sense, given that edited sites tend to be very well conserved in *Drosophila*[139,140]. Initial experimentation showed that an effective value for this bias appeared to be 4. Unless otherwise stated, this sequence conservation bias was used in all further applications of the protocol.

Given that *Drosophila* and mammals are separated by approximately 990 million years of evolution[86], it is reasonable to assume that the optimal core ECD size may differ. To investigate this we performed the protocol using a range of core ECD sizes including 15bp, 20bp, 25bp, 30bp and 40bp. Figure 5.1 (p140) shows the score distributions for the known edited exons (positives) versus all other exons (assumed negatives) for this range of core ECD sizes. Only the top intronic or exonic ECD prediction is counted for each orthologous exon pair in each graph.

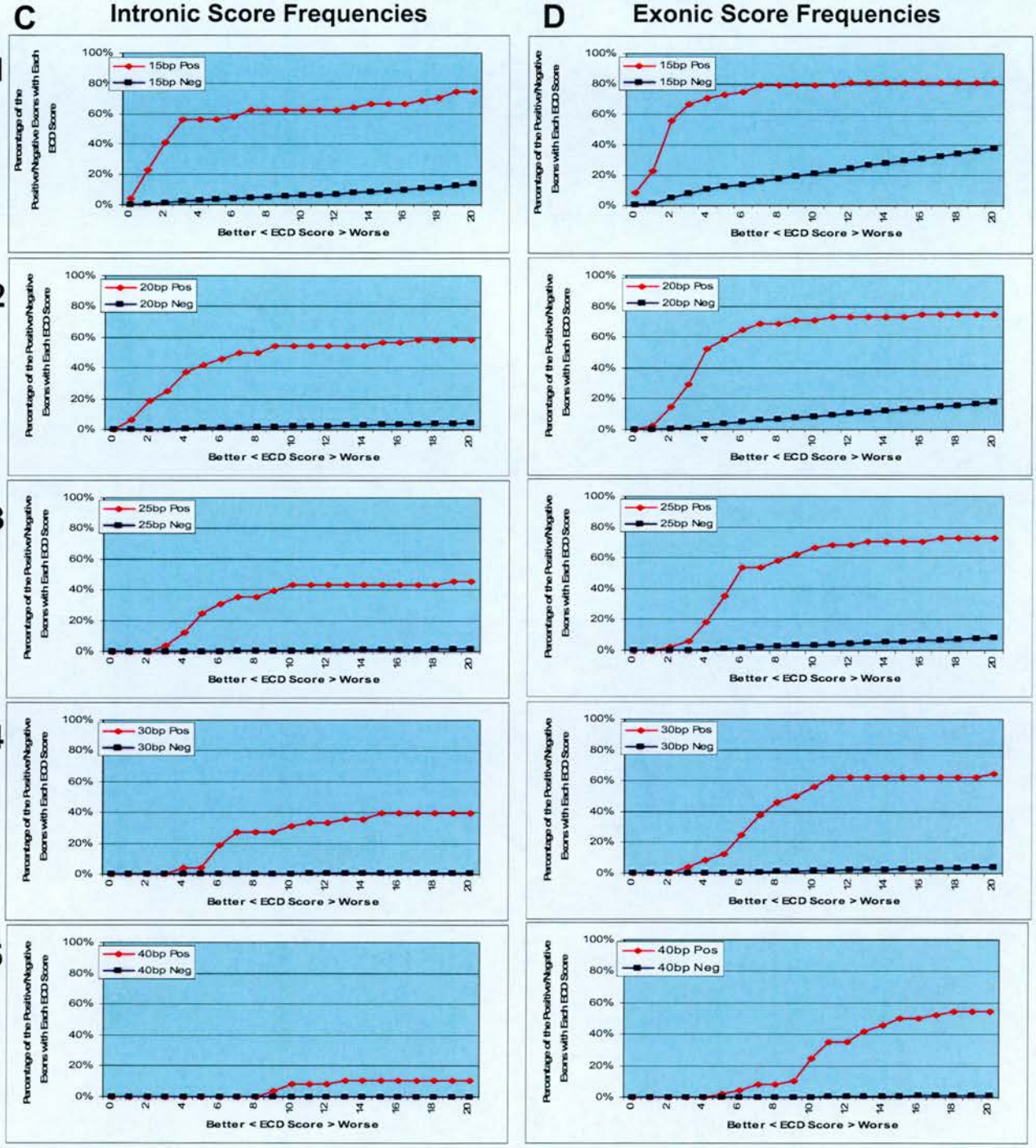
Comparing section A1 and B1 demonstrates the reason for separating the ECDs according to whether their ECSs are intronic or exonic. The background noise from

Figure 5.1. ECD Score Distributions for *Drosophila* Predictions



Discrete distributions of the intronic and exonic ECD predictions for *Drosophila*. In each case the positive ECDs (i.e. from the 48 known edited exons) and negative ECDs (i.e. from all other orthologous exon pairs (52,332)) are shown separately. The distributions do not reach 100% as some exons did not have ECD predictions that scores below 20. Half-values have been rounded up.

Figure 5.1. ECD Score Distributions for *Drosophila* Predictions (Continued)



Cumulative distributions of the intronic and exonic ECD predictions for *Drosophila*. In each case the positive ECDs (i.e. from the 48 known edited exons) and negative ECDs (i.e. from all other orthologous exon pairs (52,332)) are shown separately. The distributions do not reach 100% as some exons did not have ECD predictions that scores below 20. Half values have been rounded up.

the negative distribution is much higher for any given score in the exonic distribution. This is because ECSs that occur within exons are subject to different evolutionary forces than those that form within introns. The genuine ECS sequences in exons may be constrained by the coding sequence of the gene, possibly resulting in less perfect duplexes and worse ECD scores. Conversely, many false ECDs predictions will score very well due to the high level of sequence conservation seen in coding sequences. This makes it potentially very difficult to distinguish genuine exonic ECDs from false ones. Therefore, the scores from exonic sites are shown separately from those obtained from intronic ECDs.

There are several other observations to be made from Figure 5.1 (p140). As would be predicted, the ECD scores generally become worse as larger core ECD sizes are applied. This is observed as a shift of the distributions to the right in each graph of sections A and B. The overall sensitivity of the intronic ECD predictions also becomes worse as the core ECD size increases (as shown in sections C1-C5). The largest drop in sensitivity is between 30bp and 40bp, which have approximately 40% and 10% sensitivity, respectively. This suggests that this may be an upper cut-off for intronic ECD sizes in *Drosophila*. One possibility, as suggested by Mary O'Connell is that an upper limit may be required to ensure that there is not too much interference with/from the RNAi system. Interestingly, increasing the core ECD size does not have such a strong effect on the exonic ECD score distributions. Clearly, the scores shift to the right (see sections B1-5), but the overall sensitivity only drops from just over 80% for 15bp, to 55% for 40bp (see sections D1-5). This can be partially explained by the observation that many of these long exonic ECD predictions were palindromes. Although they are 40bp long, they fold back on themselves to make hairpins half this length. These predictions are discussed later in this chapter. These initial observations demonstrated that this method is applicable to and effective for these two *Drosophila* species.

5.5.2 Application to the Known Edited Sites in *Drosophila*

Such high levels of specificity are not required to identify the ECDs for the known edited exons as we already know these sites are edited, which necessarily reduces the potential for false positives. As such, more liberal core sizes of 15bp and 25bp were selected for intronic and exonic ECDs, respectively. Using thresholds of 3 and 6 ensured that any predicted ECDs had specificities of 97.7% and 98.4%, respectively.

Table 5.2. ECD Predictions for the Known *Drosophila* Edited Sites

Ensembl Gene	Ensembl Exon	Edited Site	Intronic ECDs Scores (15bp)		Exonic ECD (25bp) Scores	Combined Overlap
			5'	3'		
CG1522	CG1522:11	Cac K/K (12)	-	-	-	-
	CG1522:13	Cac 15	-	1	8.5	Y
	CG1522:15	Cac 17a	-	-	-	-
		Cac 17b	-	-	-	-
		Cac 17c	-	-	-	-
		Cac 17d	-	-	-	-
	CG1522:17	Cac 19	-	-	-	Y
	CG1522:21	Cac 24	-	-	-	-
	CG1522:23	Cac 26a&b	-	1	5	Y
	CG1522:27	Cac 30b	-	2	-	-
		Cac 30a	-	2	-	Y
CG9907	NULL:512413	Para t/m	-	-	-	-
	CG9907:26	Para ssp	2	2	5.5	Y
	CG9907:28	Para sfc	2	2	6	Y
	NULL:512549	Para fsp	2.5	-	-	-
CG7535	CG7535:3	GluCl 78/79	2	-	-	Y
	CG7535:8	GluCl 722	-	-	4	-
	CG7535:10	GluCl 1034	-	3	-	M
	CG7535:11	GluCl Silent 1179	-	-	-	-
CG4128	CG4128:6	Dalpha6 Cluster	3	3	-	Y
CG12598	CG12598:9	Adar	-	-	5	Y
CG18314	CG18314:5	CG18314 Clust3	-	-	-	-
		CG18314 Clust 4	-	-	-	-
		CG18314 Clust1	-	-	6	Y
		CG18314 Clust 2	-	-	-	-
CG15899	CG15899:15	CaAlpha1T	-	2	5	Y
CG4894	CG4894:5	DmCa1D Cluster	-	3	-	Y
		DmCa1D Last site	-	1	5	Y
CG12295	CG12295:5	Alpha2Delta I/Va	-	-	6	-
		Alpha2Delta I/Vb	-	-	6	Y
	CG12295:6	Alpha2Delta R/G	-	-	-	-
CG12348	NULL:538653	Shaker KERG	-	3	-	M
	NULL:538657	Shaker I/M	0	1	5	Y
	CG12348:15	Shaker I/V (sim Shab)	3	-	6	Y
		Shaker TA & QR	3	-	4	-
CG10952	CG10952:11	EAG K/Ra	-	-	-	Y
	CG10952:13	EAG Y/C	2	-	6	Y
		EAG N/D & Silent	2	-	6	M
		EAG AA	-	-	6	Y
	CG10952:14	EAG K/Rb	-	-	5	Y
CG10693	CG10693:6	Slowpoke N/D	1	3	4	Y
	NULL:1763569	Slowpoke S/G	3	-	5	Y
CG3139	NULL:915362	Syt I/V - A	2	1	-	Y
		Syt K/R & I/V - B/C	1	2	-	Y
		Syt I/M - D	1	3	-	Y
CG2999	NULL:2361787	Unc-13 S/G	-	-	6	M
CG32490	CG32490:6	Cpx I/M NDSG	2	-	6	Y
CG2520	CG2520:10	Lap T/A	-	-	-	-
CG32975	CG32975:5	Da5 I/V (Nic34E)	2	2	5.5	Y
	CG32975:6	Da5 T/A I/V L/L (Nic34E)	-	1	5	Y
CG11348	CG11348:3	ARD R/G 64b	1	3	6	Y
	CG11348:4	ARD I/M 64b	-	-	3	Y
CG6798	CG6798:7	SBD T/A & Silent (96A)	2	-	-	Y
CG10537	CG10537:4	Rdl L/L R/G	-	-	-	-
	CG10537:7	Rdl I/V	3	0	5	Y
		Rdl N/D	3	0	5	Y
	CG10537:8	Rdl M/V	1	-	-	Y
CG1066	NULL:368787	Shab I/V	-	-	-	-
CG13167	CG13167:1	CG13167 I/V	-	-	-	-
CG9619	CG9619:3	CG9619 S/G	-	-	-	-
CG8428	NULL:2096895	CG8428 N/G Spin	-	3	-	Y
CG12076	CG12076:3	CG12076 Q/R	-	-	-	-
CG14936	CG14936:3	CG14936 V/V	-	-	-	-
CG14936	NULL:1219665	CG14936 G/G	-	-	-	-
26 Genes	48 Exons	64 Sites/Clusters	12 Overlap 5'prime	13 Overlap 3'prime	16 Overlap (Exonic)	35 Combined Overlap

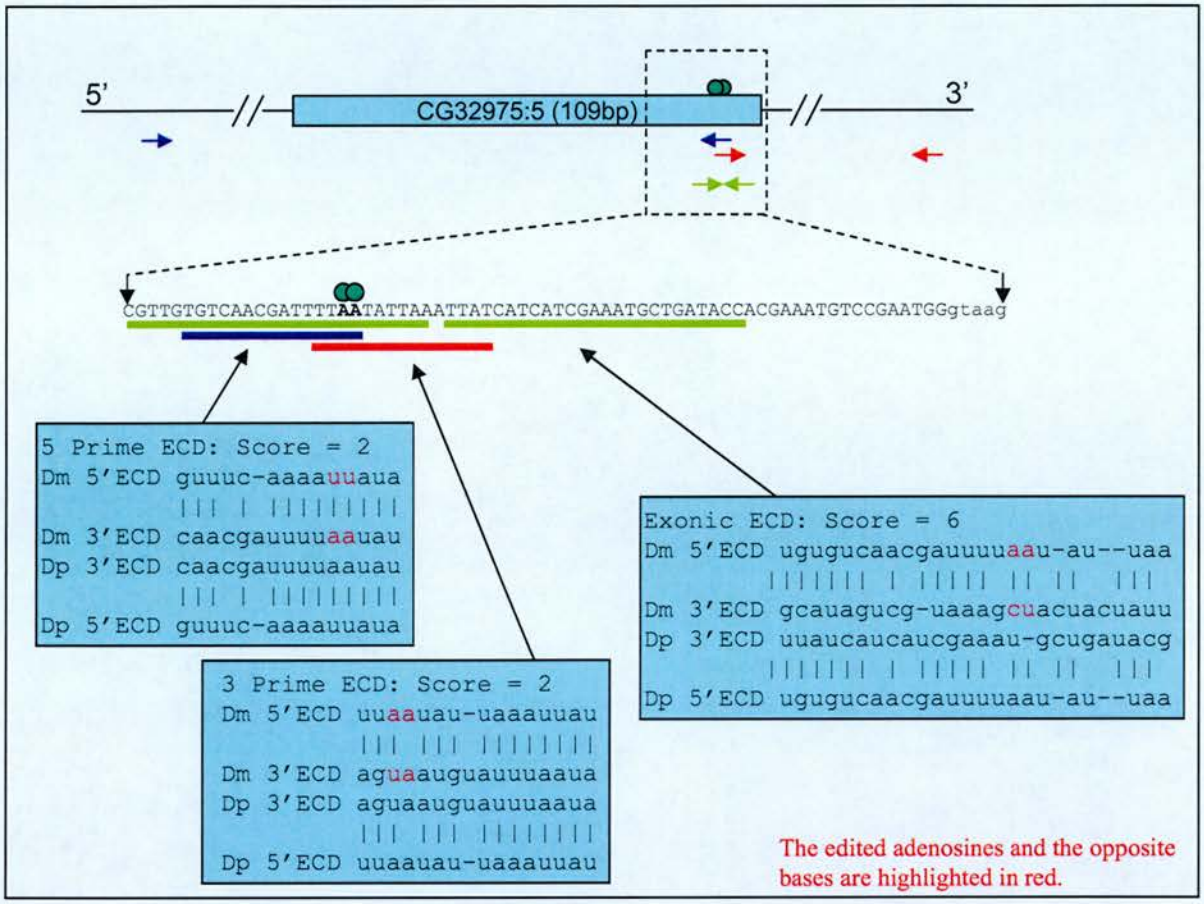
Legend. The Known *Drosopila* edited sites are arranged per gene, per exon and per cluster. A cluster is considered to be any group of edits that could occur within the same hairpin (of an assumed length of 30bp). The ECD predictions scores are shown in the adjacent columns. The 5'prime intronic & 3'prime intronic scores are derived using a core ECD size of 15bp and a conservation bias of 4. Only intronic ECDs scoring 3 or less are shown in the table. These results have a specificity of 97.7%. The exonic ECD prediction scores are derived using a 25bp core EC size and a conservation bias of 4. Only exonic ECDs scoring 6 or less are shown in the table. These results have a specificity of 98.4%. StnB had no orthologous exon and was removed. The fin column describes which editing sites have at least one overlapping ECS prediction (Y), or at least one ECS prediction within 15bp (M).

Table 5.2 (p143) shows the results for the known edited sites in *Drosophila.* Forty-two of the 64 sites/clusters have an ECD prediction within these strict thresholds. Only five of these sites/clusters have ECD predictions that are more than 15bp from the individual edited sites. The remainder have ECD predictions that overlap(33) or are within 15bp of the edited sites(4). These are shown in red and orange, respectively. In 33 out of 37 of these cases, the overlapping ECD is the best scoring ECD for that exon (scores are shown in bold). These results demonstrate that this protocol can sensitively and specifically identify ECDs that overlap edited sites.

Table 5.2 (p143) also demonstrates that many exons contain more than one high quality (i.e. low-scoring) ECD predictions. This was not expected due to the very high specificities of these analyses. For example, the *Da5* T/A site has an exonic ECD overlapping it, but there is also a 5 prime ECD prediction with a score of 1 located elsewhere in the exon. It is possible that these predictions overlap uncharacterised edited sites. It is also possible that the high levels of sequence conservation often seen in edited exons could have resulted in random ECD predictions of high quality. In many cases there were two or more high quality ECD predictions that both overlapped the edited nucleotides. For example, Figure 5.2 (p145) shows three ECD predictions that overlap the *Da5* I/V site. The putative structures formed by each of these ECDs would block the formation of the other two ECDs. It is possible that all three of these ECD predictions are functional and that they can all form *in vivo,* albeit not simultaneously. The result of this could be that each ECD blocks or enhances the editing of the two nucleotides to differing degrees. Variation in the opposite base[17] and the location of bulges in the duplex[77] have been shown to affect the editing efficiencies. The three ECDs shown here each have different duplex structures, with different bases opposite the edited adenosines in each case, suggesting that the ECDs could indeed result in different editing efficiencies. This model could allow for greater control over when and where particular bases are edited. However, without carrying out laborious experimental work, it would not be possible to confirm these possibilities.

In contrast to mammalian ECDs, where the majority occur in the 3 prime introns, the *Drosophila* sites displayed strong ECDs in both introns and in exonic sequences. It was unclear whether these trends underlie a difference in the function or evolution of ECDs, or whether the imbalance is stochastic. There were many other cases, however, where there were no ECD predictions overlapping the edited nucleotide.

Figure 5.2. Multiple ECD Predictions Overlapping the Da5 I/V Edited Site



Three ECD predictions are shown for the Da5 I/V site. These include a 5 prime intronic ECD, a 3 prime intronic ECD and an exonic ECD. The edited sequence is expanded and the putative ECD structures are displayed. Each ECD is split into it's 5 prime and 3 prime halves. The edited adenosines and the opposite bases are highlighted in red. The opposite bases and the location of potential bulges vary between the three ECD predictions.

The possible reasons for this are described in the next section.

5.5.3 Why Are Some ECDs Missing?

Firstly, the ECD may be conserved, but the degree of sequence conservation and the quality of the secondary structure may be too poor to identify using this protocol. The secondary structure could also be too complex to identify as a single duplex. For example, the structure of the *syt* (*synaptotagmin-I*) B,C&D sites has been proposed to be a long range pseudo-knot[108]. It is not clear how common these pseudo-knot structures will be in edited sites, given that only one example has been identified so far.

Alternatively, the ECDs may be missed because of the parameters of the protocol. For example, if the ECD is much shorter than 15bp, for intronic ECDs, or 25bp, for exonic ECDs, then it is unlikely to be identified. Alternatively, if the second half of the ECD is located more than 2kb away from the exon ends, then it will not be found. Additionally, this protocol is restricted to find ECDs only within the exon and neighbouring introns. It is possible that the other half the ECDs is located in more distant introns or exons. In order to test these possibilities, the protocol was repeated with the distance restraints relaxed so that 5kb on either side of the exon was scanned for ECDs, even if this extended into neighbouring exons or more distant introns. This analysis did not identify any additional ECD predictions that overlap the known edited sites and were below the previous thresholds. The only exception to this was the observation of an additional ECD prediction for the *syt* D edited site. This suggests that most of the ECSs for *Drosophila* edited sites occur within 2kb of their edited sites.

Finally, it is possible that the editing of these sites is not mediated by an RNA duplex composed of two sequences in the same transcript (i.e. *in cis*). Given that dADAR contains a double stranded RNA binding domain, this could suggest that a duplex is being formed *in trans*. Although there is absolutely no evidence for this it is important to state the possibility. If, however, we assume that these sites are edited via an ECD, then it is possible that the ECD is not conserved between *D.mel* and *D.psu*. This could indicate that these exons are not edited in *D.psu* or that they are mediated by a different ECD. If all the missing ECDs were due to a lack of conservation, then this could indicate rapid rates of emergence and/or loss for ECD

structures in *Drosophila*.

5.5.4 The Published ECDs

Apart from sensitivity and specificity estimates, another way to gauge the success of this protocol is to compare my predictions to the published ECD predictions. Table 5.1 (p132) shows the seven edited sites that have ECD predictions published. However, only the *syt* B,C&D sites have an experimentally validated structure. The published structure for the *syt* B&C sites agrees with our ECD prediction within the 3 prime intron (score was 2). A five prime ECD prediction was also identified for this site. The *syt* D site has two 3 prime ECD predictions overlapping the edited site. The second ECD prediction, which was identified using the extended scan of 5kb of intron sequence, agreed with the published ECS. This shows that some functional ECSs *can* be located over 2kb from the edited site.

The other published ECD predictions have not been validated. The ECD prediction for the *para* ssp site agreed with the published prediction, but no matching ECD predictions were found for the other four sites. It is not clear whether the published predictions are wrong, our protocol predictions are wrong, both are wrong or that both our predictions and their predictions are functional. Even if the ECD predictions are not strictly correct, the protocol is still useful, as it remains highly specific for identifying edited exons.

5.5.5 Exonic Palindromes

Figure 5.1 (p140), sections D1-5 demonstrate that there was a high proportion of the known edited exons that had long exonic ECD predictions. For example, 46% of the known edited exons have 40bp exonic ECDs scoring less than or equal to 14. In comparison, less than 0.7% of other exons have 40bp exonic ECDs of the same quality.

However, many of the sequences in the predicted ECD structures were palindromes (e.g. the *cac26* site). In other words, the ES and ECS are located in the same part of the sequence, but in opposite directions. This is a problem with searching the exon sequence against itself. In these cases, although the core ECD size used to predict

these sites was 40bp, the actual duplexes, which are hairpins, can only be half this length. For example, the sequence AAAAUUUU would be recognised as base-pairing against its reverse to create an 8bp duplex. However, this is not physically possible in *cis* and the longest duplex would be 4bp (all the 'A's base-paired to all the 'U's). In reality, the duplex would be even shorter as several nucleotides are required for the RNA to fold back on itself. This means that many of these supposed 40bp ECDs are actually only 20bp long. However, some of the 40bp ECD predictions do not overlap, or do not overlap completely, resulting in longer duplex predictions. For example, the 25bp exonic ECD shown in Figure 5.2 (p145) can also be seen as a slightly palindromic 40bp ECD prediction. In reality, however, the duplex shown could not be greater than 28bp. Although these ECDs are actually shorter than previously suggested, they are still predicted with very high specificity (99.3%), which suggests that they are still as likely to be functional.

5.5.6 Summary

In summary, this protocol has been successful in identifying high quality ECD predictions for over half of the known edited sites. It is not possible to ascertain if all of these are the genuine ECD structures, however, the only two ECD structures that have been experimentally validated are both identified using this protocol.

5.6 A *Drosophila* Genome ECD Screen

The thresholds used to identify predicted ECDs for the known edited exons were not considered strict enough for a full genome screen. Based on these graphs in Figure 5.1 (p140), it was decided that the optimal ECD size for a genome scan for intronic ECDs was 20bp. The proposed optimal ECD size for a genome scan for exonic ECSs was 40bp. These sizes were chosen as they represented a compromise of high levels of sensitivity with very high levels of specificity. For example, using threshold scores of 4 for intronic ECDs and 14 for exonic ECDs provided sensitivities of 38% and 46%, with specificities of 99.44% and 99.34%, respectively. Combining these analyses resulted in an overall sensitivity of 58% and a specificity of 98.9%. These high levels of specificity are required due to the large number of putative orthologue pairs in the genome. These three specificities represent 292, 344 and 587 false positives, respectively (out of a total of 52,332 orthologous exon pairs). In reality some of these are likely to be genuine. The results and analyses of these screens are presented here.

5.6.1 A Screen for Intronic ECDs

Using a core ECD size of 20bp and a threshold ECD score of 4 or less, results in obtaining 18 of the known edited exons (38%), with an additional 292 exons (0.56%). This represented far too many putative edited exons to test experimentally. However, simple observation of the known edited exons shows that 34 of the 48 exons have another edited exon in the same gene. Based on this, we applied a filter to these results such that only genes with more than one exon containing an ECD scoring less than or equal to 4 were considered. The results of this analysis are shown in Table 5.3 (p150). Of the 18 known exons with ECDs less than or equal to 4, twelve passed this filter (as described at the bottom of Table 5.3). In contrast, only 44 putatively novel edited exons passed the filter. This represents a sensitivity of 25% and specificity of 99.92%.

Only six of the ESs from these novel ECD predictions overlap non-coding sequences, repeats, or simple sequences (e.g. AT rich sequences). The remainder represent a relatively high quality set of ECD predictions. These predictions are roughly equally split between ECDs with 5 prime and 3 prime ECSs (21 and 17

Table 5.3. A Genome Screen for Novel 20bp Intronic ECD Predictions

Ensebml Gene	Ensembl Exon	ECD Score	Where	Coding	Simple Repeats	Brief Description and Notes for Each Gene (Collected from Ensembl & FlyBase)	Potential Groups?	Over Exon Boundary
CG17117	CG17117:7	1	3prime	Yes	No	Homothorax homeoprotein. Uncoordinated in C.Elegans. Homeobox myeloid ecotropic viral insertion site-2a protein. Brain development. Expressed in embryo and larva.		No
	CG17117:6	3.5	3prime	Yes	No			Yes
	CG17117:16	3.5	3prime	No	AT rich			Yes
CG7499	CG7499:1	2	3prime	Yes	No	CG7499. Rh50-like protein. Rhesus blood group-associated B glycoprotein.		Yes
	NULL:404733	4	5prime	Yes	No			Yes
	NULL:404735	4	3prime	Yes	No			No
CG32206	CG32206:5	2	5prime	Yes	AT rich	LDL Receptor.		No
	CG32206:11	4	5prime	Yes	No			Yes
CG9995	CG9995:11	2	3prime	Yes	No	Huntingtin. HEAT repeat. Zinc carboxypeptidase A metalloprotease. Gene mutated in Huntingdon's Disease. Possibly also occurs in mouse & rat. Drosophila mutant has axonal transport problems, then develops rough eye and severe problems with morphology and pigmentation, leading to dark areas indicative of death.	Axon G.	Yes
	CG9995:14	4	5prime	Yes	No			No
CG31298	CG31298:4	3	5prime	Yes	No	Beat Vb (Fragment). Beaten path. Immunoglobulin proteins involved in axon guidance.	Axon G.	Yes
	CG31298:3	4	5prime	Yes	No			Yes
CG32922	CG32922:2-A	3	5prime	Yes	No	Bicistronic with gene called Skeletor (involved in mitotic matrix). DOMON domain. Dopamine / catecholamine metabolism.	Receptor / Channel	Yes
	CG32922:1-A	4	3prime	No	No			No
CG10443	CG10443:12	3	5prime	Yes	No	Protein- tyrosine-phosphate phosphohydrolase (dLAR). Immunoglobulin axon guidance. Embryonically expressed. Lots more data.	Axon G.	No
	CG10443:4	4	3prime	Yes	No			Yes
CG1470	CG1470:15	3	3prime	Yes	No	Soluble guanylyl cyclase beta subunit. Embryonic and larval expression.		Yes
	CG1470:3	3	5prime	Yes	No			No
CG17800	CG17800:17	3	3prime	Yes	No	Human homologue is Down syndrome cell adhesion molecule like 1. Involved in axon guidance & peripheral nervous system development (esp. mushroom body).	Axon G.	No
	CG17800:18	3.5	5prime	Yes	No			No
CG9652	CG9652:2	3	5prime	Yes	No	Dopamine receptor 1 precursor (D-DOP1). Lots of different receptor domains.	Receptor / Channel	No
	CG9652:5	4	3prime	Yes	No			Yes
CG6963	CG6963:11	3	3prime	Yes	No	Gish (gilgamesh). Casein Kinase. Involved in spermatogenesis & glia cell migration. Neuronal development in the eye.	Axon G.?	Yes
	CG6963:3-A	3	3prime	Yes	No			Yes
CG11020	CG11020:13	3	5prime	Yes	AT rich	Mechanosensory transduction channel NOMPC. Ion transport. Expressed in pupa. Various morphological phenotypes. Especially sound responses.	Receptor / Channel	Yes
	CG11020:16	4	5prime	Yes	No			No
CG31187	CG31187:10	3	5prime	Yes	No	Diacylglycerol kinase.		Yes
	CG31187:7	3	3prime	Yes	No			No
CG7147	CG7147:3	3.5	5prime	Yes	No	Kuzbanian. Metalloproteinase/disintigrin involved in Notch signalling, CNS development and regulation of axon extension. Expressed in embryo and larva.	Axon G.	Yes
	CG7147:13	4	3prime	Yes	No			Yes
CG12478	CG12478:4	3.5	3prime	Yes	No	Homologous with human trinucleotide repeat containing 4. RNA-binding orphan nuclear receptor. Paraneoplastic encephalomyelitis antigen. Called Bruno-3. May have/be a miRNA.		Yes
	CG12478:11	4	5prime	Yes	No			Yes
CG5403	NULL:257661	3.5	5prime	Yes	No	CG5403. Dead ringer protein (Retained protein). Transcription factor with various functions. It's required for positioning of the longitudinal glia in the embryonic CNS.	Axon G.?	Yes
	NULL:257668	4	5prime	Yes	No			No
CG10738	CG10738:11	4	3prime	Yes	No	Uncharacterised gene. GABAB/ANF receptor? Guanylate cyclase? Tyrosine kinase?	Receptor / Channel	No
	CG10738:3	4	3prime	Yes	No			No
CG31163	CG31163:18	4	5prime	Yes	No	Human SH3 protein expressed in lymphocytes homolog.		No

Table 5.3. A Genome Screen for Novel 20bp Intronic ECD Predictions

Ensebml Gene	Ensembl Exon	ECD Score	Where	Coding	Simple Repeats	Brief Description and Notes for Each Gene (Collected from Ensembl & FlyBase)	Potential Groups?	Over Exon Boundary
	CG31163:20	4	3prime	No	AT rich			Yes
CG7178	CG7178:10	4	5prime	Yes	No	Troponin I (TNI) (Wings apart-A protein) (Heldup protein). Actin binding, neurogenesis, muscle development, anti-freeze domain. Similar to synaptotagmin? Expressed in embryo & larval stages. Mutations involved in cuticle, nervous system and muscle development.		Yes
	CG7178:11	4	3prime	Yes	No			Yes
CG8442	CG8442:1	4	5prime	Yes	No	Glutamate receptor I precursor (dGLUR-I) (Kainate-selective glutamate receptor). Expressed in embryo.	Receptor / Channel	Yes
	CG8442:5	4	5prime	Yes	No			Yes
CG14064	CG14064:5	4	5prime	Yes	No	Beat VI. Immunoglobulin domain. Beaten path V1. Immunoglobulin proteins involved in axon guidance.	Axon G.	No
	CG14064:2	4	5prime	No	No			No

This includes the following known edited exons: Cac(15, 26), Da5(I/V, T/A), Para (fsp, ssp, sfc, fsp), Rdl(I/V, M/V) & Shaker(KERG, I/M).

LEGEND: Table 5.3 shows the novel *Drosophila* intronic ECD predictions. They are arranged per gene and are annotated with the best intronic ECD score per exon, the location (5prime or 3prime) of the ECS, the coding potential of the sequence, and whether the ECD contains repeats or simple sequence. Gene descriptions have been compiled from the Ensembl and FlyBase databases, including notes on the gene phenotypes where available. The penultimate column annotated genes that appear to fit in the following categories; Axon Guidance (demonstrated invovlement in the relocation of neural cells), & Receptor/Channel (gene appears to be a channel or receptor). The final column shows whether the ES overlaps the exon boundary.

respectively). More interestingly, however, 26 of these ECD predictions overlap one of the boundaries of their exons. As discussed in the previous chapter, these predicted structures could function in the regulation of splicing, editing, or both.

The known *Drosophila* edited sites are primarily involved in the nervous system[34]. This is reflected by the novel predicted ECDs, many of which appear to be in transcripts for channels or receptors (see Table 5.3 – p150). These include two dopamine receptors, a mechanosensory channel, a putative GABA receptor and an ionotropic glutamate receptor. Ensembl does not predict the glutamate receptor to be orthologous to any of the known edited mammalian glutamate receptors. There are also 6 genes that appear to be involved in axon guidance or axon transport. Four genes are directly associated with axon guidance, including two *Beat* genes, the *DsCam* gene and a protein tyrosine-phosphate phosphohydrolase. Additionally, the *Huntingtin* gene is involved in axonal transport and the *Kuzbanian* gene is involved in axon extension. There are also two genes in this list that are involved in glial cell positioning. These are *Gish*, a casein kinase, and *Dead ringer*, a transcription factor. There was no established GO term for axon guidance, so it was not possible to quantify the significance of these results. However, these observations make a reasonably strong case for the role of editing in axonal guidance and function in the fruit fly.

The *DsCam* gene has been previously mentioned in this thesis as it contains a complex example of conserved RNA duplexes controlling alternative splicing. The duplexes we identified did not appear to be related to those described by Graveley *et al*[120]. Some of the results of this screen are being experimentally tested for editing, although no results are currently available.

5.6.2 A Screen for Exonic ECDs

Using a core ECD size of 40bp and a threshold ECD score of 14 or less, results in ECD predictions for 22 of the 48 known edited exons and 344 of the 52,332 other exons with putative orthologues in *D.psu*. Again, this represented far too many putative edited exons to test experimentally. The filter applied to the intronic ECD results did not sufficiently reduce the number of exonic ECD predictions (i.e. more than one exon per gene with an ECD below the threshold). Instead, the number was reduced by applying a stricter threshold score of 8 or less, but only requiring one

exon per gene. This resulted in ECD predictions for 4 of the known edited exons and 24 additional exons, which represents a more manageable number of predictions.

The top 24 exons are from 24 genes and are described in Table 5.4 (p154). These include *eIF4A*, two calcium channels (*Sk* & *CG6320*), a putative potassium channel (*CG9817*), two orphan nuclear receptors (*Bru-3* & *Akap200*), a GABA-B receptor, a splicing gene (*CG31550*), a synaptic gene (*Tomosyn*), and 15 other genes. Sixteen of these exonic ECD predictions overlap coding sequences. Eight of the ECDs comprise AT rich sequences. More interestingly, only one of these ECDs is not a palindrome, or a partial palindrome. This means that most of these structures are approximately 20bp long. Four of these genes have A-G mismatches that can be observed nearby in the publicly available cDNA sequences. Five of the six cDNAs containing the mismatches near these four ECD predictions are from 0-24hour *Drosophila* embryos. These ECDs are currently being experimentally tested. The *eIF4A* gene is the best candidate as it has the best ECD score and has an associated A-G mismatch. It was also identified as the 84th most enriched gene in the Xia *et al* inosine antibody-based screen[133]. The *Kuzbanian* gene was identified in both the intronic and exonic screens, suggesting that this is also a strong candidate. Equally, the exonic screen contains another *Beat* gene (*Beaten path Ic*), in addition to the two identified in the intronic screen. One of the top candidates in the vertebrate screen was *Neuroligin 3*, which has been shown to interact with neurexins, of which three ranked 100th, 101st and 106th. This exonic screen contains a neurexin (*Nrx-1*). Together, these observations suggest that there might be conservation of editing in neurexins between *Drosophila* and vertebrates.

5.6.3 Comparison to External Data

In addition to the known edited sites, there were two external datasets that our data can be usefully compared to. Xia *et al* used an inosine antibody to identify transcripts containing inosines. They successfully validated editing in only six novel transcripts, but predictions were made for a much larger number of transcripts[133]. We compared this list of putative inosine-containing transcripts with our list of putative ECDs. Unfortunately, the Xia list only identified one gene from either the intronic or exonic screens. This was *eIF4A*, the best candidate from the exonic screen. This result is not entirely surprising as only one of the previously known edited transcripts was in this list (CG18314).

Table 5.4. A Genome Screen for Novel 40bp Exonic ECD Predictions

Ensembl Gene	Ensembl Exon	ECD Score	Brief Description of Gene	AG overlap	Coding	Simple Repeats	Palind-romic
CG9075	CG9075:5	6	Eukaryotic initiation factor 4A (eIF4A). Larval development. 86th in Xia *et al* inosine pulldown.	1(+8)	Yes	No	Yes
CG9297	CG9297:1	7.5	Dynamin/thrombospondin containing protein?	1(+5)	No	No	Partial
CG6821	CG6821:1	8	Larval serum protein 1 gamma chain precursor (Hexamerin 1 gamma). Hemocyanin	1(+15)	Near	No	Yes
CG9821	CG9821:1	8	Putative phosphoglycerate kinase	1(+10), 1(-10)	No	AT rich	Partial
CG1028	CG1028:10	6	Homeotic antennapedia protein	0	No	AT rich	Yes
CG10706	CG10706:20	6	Calcium-activated SK potassium channel and Calmodulin binding domain containing protein family member	0	Yes	No	Yes
CG11328	CG11328:4	6	Nhe3. Sodium/hydrogen exchanger	0	Yes	No	Yes
CG31550	CG31550:3	6	Splicing factor 4 family	0	Yes	AT rich	Yes
CG32045	CG32045:23	6	Fry - Wing/bristle morphogenesis	0	Yes	No	Yes
CG9817	CG9817:4	6	Uncharacterised gene with following domains: Ribosomal protein P2, Involucrin, Eggshell, Antifreeze, Kv1.1 channel.	0	Yes	AT rich	Yes
CG4838	CG4838:5	6.5	Beaten path Ic	0	Yes	No	Partial
CG13620	CG13620:2	7	Uncharacterised gene with following domains: Involucrin, Zn finger	0	Yes	No	Yes
CG31044	CG31044:1	7	Uncharacterised gene with antifreeze domain	0	No	No	No
CG7050	CG7050:11	7	Nrx-1. Neurexin (cell adhesion)	0	Yes	No	Yes
CG10601	CG10601:1	8	Mirror (Homeoprotein Sail). PNS development	0	Yes	No	Yes
CG12478	CG12478:2	8	Bruno-3 - Orphan nuclear receptor	0	Yes	No	Yes
CG13388	CG13388:6	8	Akap200 - Orpah nuclear receptor	0	No	AT rich	Partial
CG15274	CG15274:2	8	Metabotropic GABA-B receptor subtype 1	0	Yes	No	Yes
CG15275	CG15275:2	8	Uncharacterised gene with Immunoglobulin-like domain	0	Yes	No	Yes
CG6320	CG6320:13	8	Voltage-dependent L-type calcium channel beta subunit	0	Yes	No	Yes
CG6803	CG6803:1	8	Possibly similar to myofibril-associated Zeelin1 protein	0	No	AT rich	Partial
CG7147	CG7147:13	8	Kuzbanian	0	No	Very AT rich	Yes
CG7467	CG7467:9	8	Osa. Trithorax group protein OSA (Eyelid protein). Antifreeze. Involucrin. Eye development and much more.	0	Yes	AT rich	Yes
CG17762	NULL:445683	8	Tomosyn. Involved in synaptic release (similar to synaptobrevin). Involucrin. Lethal to giant larvae. Cell polarity.	0	Yes	No	Yes
This includes the following known edited exons: Da5 (T/A), ARD (I/M), Para (sfc) & Shaker (I/M)							
LEGEND: Novel exonic ECD predictions that scored 8 or less. The top four sites have A-G mismatches nearby (distance in brackets).							

In contrast, Glazov *et al* performed an analysis to identify the best conserved elements, which they termed ultra-conserved, between these two species of *Drosophila*[88]. This data gives a more considerable overlap with our data. In particular, Glazov *et al* present a table of the 10 genes harbouring the largest ultra-conserved sequences between *D.mel* and *D.psu* overlapping exons and splice sites. This list contained six of the known edited genes and two of my candidate edited genes, one from each screen. These two genes were *Homothorax,* a homeobox protein, and *Sk*, a calcium-activated potassium channel, from the intronic and exonic screens respectively. Glazov *et al* also noted that *Homothorax* forms a duplex, which prompted them to test this site for editing. They did not identify any sign of editing in the samples they tested (mixed stage embryos). They also identified the *Bruno-3* gene, which was identified in our exonic ECD screen, as containing over 11kb of ultra-conserved sequence in 168 elements[88]. The reasons for such extensive conservation remain unknown.

5.7 Summary

This chapter contains a variety of ECD predictions, both for known edited sites and novel candidate exons. Given the performance of these methods on the known edited sites and the two known ECD structures, it seems likely that some of these novel predictions will be genuine and they may provide explanations for the presence of 'ultra-conserved' non-coding elements.

6 Results: Online ECD Prediction Tool

6.1 Preface

Chapters 4 and 5 demonstrate the application of a protocol to identify ECDs conserved between two species. In order to make this method applicable to other species and accessible to other users, it has been mounted on a web server. This allows future potential users to search any specified sequences for conserved ECDs.

This resource could be used to look for conserved ECDs in other species. One future project could use this to look for conserved ECDs between two nematode species, or two fish species. These would be the first computational analyses for A-I editing based in these species. This resource could also be used to identify the known ECDs in additional species to those analysed in this thesis.

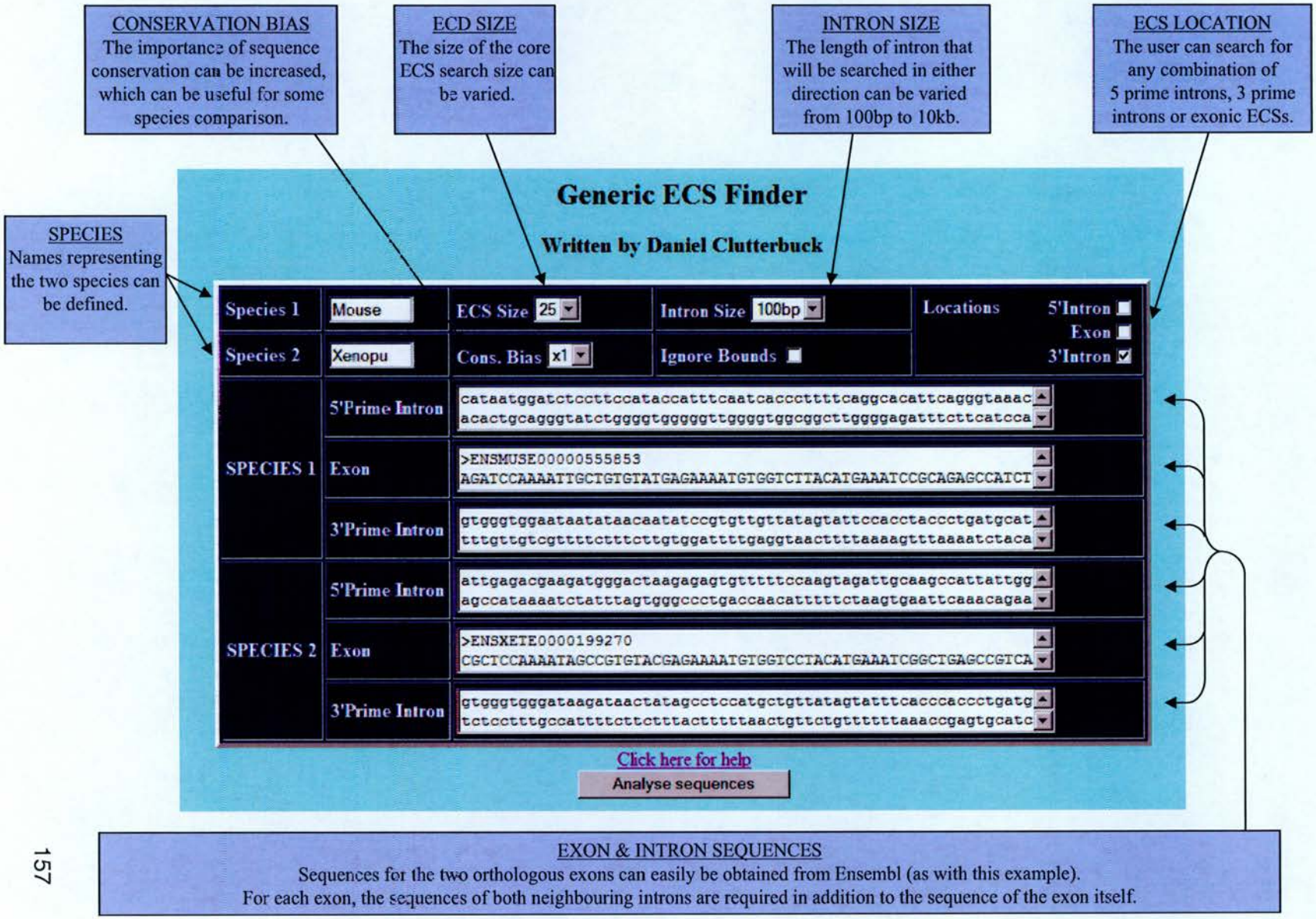
6.2 Materials & Methods

The method underlying this resource uses a combination of an HTML input form, CGI, PHP and a Perl script. This Perl script is based on the main scripts used in the previous two chapters. The other components are standard web design features to allow communication with the user and do not need to be discussed in any detail.

Figure 6.1 (p157) shows the HTML input form for this resource. The user must provide the following information;

- Species names for the two species to be analysed.
- Core ECD Size (as defined in Chapter 4).
- Conservation bias (as defined in Chapter 5).
- Intron Size (as defined in Chapter 4).
- ECS Locations (can be any combination of 5 prime intron, 3 prime intron and exon).
- 5 prime intron, 3 prime intron and exon sequences for the two species being analysed. These can easily be obtained from Ensembl.

Figure 6.1. The HTML Front End of the Online ECD Prediction Tool



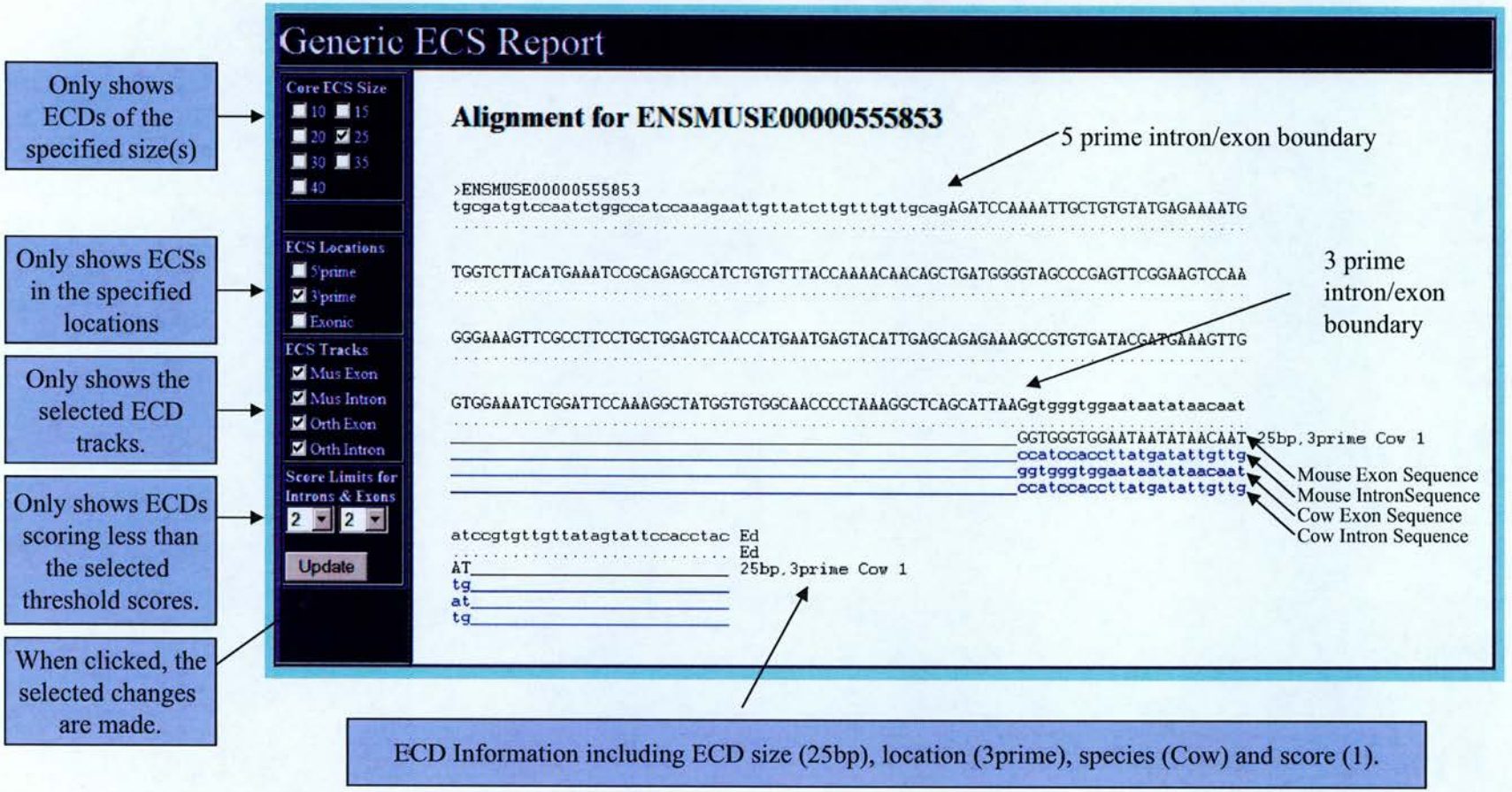
When the form is completed, the various scripts are run and an HTML results page is returned. An example is shown in Figure 6.2 (p159). This displays the exon sequence, with an extra 50bp flanking sequence on either side. The ECDs are displayed with the ES of the first species aligned to this sequence. The output is dynamic so that only ECDs meeting strict user-defined criteria are shown. This feature has been achieved using Javascript. These criteria include the ECD core size, the ECS locations, which ECD sequences to display and ECD score thresholds for both intronic and exonic ECDs.

6.3 An Example Application

Figure 6.2 (p159) shows the results of applying this resource to identify ECDs between the mouse exon containing the R/G site and its orthologous exon in cow. The cow ECD for this exon has not been published, but based upon conservation across other mammals, it should be detected by this protocol. The output has been restricted to display only 25bp ECDs, with ECSs in the 3prime intron and ECD scores of 2 or less. Using these criteria, the only ECD shown is the published ECD for the mouse *GluR-C* R/G site and its orthologous ECD in cow.

This resource is undergoing testing and may be published and made publicly available at a later date. A command line version of the program may also be made available. This program is included in the attached CD.

Figure 6.2. The Dynamic HTML Output of the Online ECD Prediction Tool



7 Discussion

7.1 Preface

There were two primary aims for this thesis, which were to improve the understanding of the known edited sites and to identify novel, or potentially novel edited sites. Both of these aims have been achieved.

7.2 Mismatch-Based Screen for A-I Editing

Chapter 3 describes a protocol for using A-G mismatches to identify putative edited sites within coding sequences in the mouse and human genomes. Seven features, based on the known protein recoding edited sites, were used to identify and rank novel candidate edited sites. The top 10 sites were experimentally tested and the top scoring candidate, *BC10*, was shown to contain genuine edited sites[79]. A similar paper, which was published while our work was under review, also identified this edited site[81]. Together, these papers were the first bioinformatic approaches to identify more than one of the known mammalian sites, or to successfully identify any novel protein recoding sites. Based on the number of sensitivity and specificity of our protocol, we were able to make the first statistically derived estimate of the total number of protein recoding A-I editing sites in mammals. This suggested that recoding sites are relatively rare, and that only a few more sites remain undiscovered. However, there are a number of caveats to this calculation, which have been discussed in Chapter 3. Several of the known edited sites were not picked up by this protocol, as there were no A-G mismatches in the publicly available expressed sequences at these sites. In addition, this protocol was only able to identify edited sites within Ensembl exons. Since this work was published, it has become clear that there is a link between A-I editing and Amyotrophic Lateral Sclerosis[187]. The fifth best candidate in this screen was called '*ALS 2 Chromosome Region Candidate 9*', which suggests that editing of this gene may be involved the aetiology of this disease. This link is currently under investigation.

7.3 Conserved ECD Screen in Vertebrates

One component of the protocol in Chapter 3 was the identification of putative ECDs. This preliminary method appeared to be the first published method for rapidly identifying putative RNA duplexes. A number of more complex programs are available, but these are all very slow when dealing with large sequences[175]. The protocol in Chapter 4 built on this preliminary method, primarily by changing it from a search for ECDs within a single species, to a search for ECDs *conserved* between two species. As this method does not require A-G mismatches, it can be applied to any pair of species for which orthologous sequences are available. Although this work has not yet been published, it would represent the first method for identifying ECDs. A statistical method was designed, such that ECD predictions for all the exons in the mouse genome could be scored and ranked, based on the conservation of their sequences and structures with five other species. These species were rat, human, chicken, zebrafish and pufferfish. This method was able to identify the correct ECDs for 9 of the 10 known recoding edited sites with published ECDs, demonstrating that it is sensitive. A convincing ECD structure was also predicted for the *GluR-6* I/V site, which has not been previously reported.

This method was applied to all mouse exons with orthologous exon predictions in any of the other species listed above, which incorporated over 200,000 exons. The 6 top scoring exons included 4 known edited sites, demonstrating that this method can also be highly specific. Applying a suitable threshold (combined LOD score > 17.5) resulted in a sensitivity of 60% and a specificity of almost 100%. Interestingly, the exons that ranked 2nd, 3rd, 7th and 8th were a homologous group of exons from AMPA receptors. The location of the predicted ECD suggests that it would overlap these exons' branch sites and could be involved in the regulation of alternative splicing. Alternative splicing of this exon, termed the 'flip' exon, has been shown to have a major effect on the gating properties of the resultant AMPA receptors[33]. In addition to these sites, there were many more candidates, some of which are being experimentally tested. Nineteen of these candidates had more than one exon per gene in the top 100. This observation is in agreement with the fact that many of the known edited genes contain more than one edited exon, and suggests that these candidates are more likely to be genuine. Some of these candidates are being experimentally tested. There are a number of caveats associated with this protocol, which are discussed in Chapter 4.

In retrospect, there are several improvements that could be made to this protocol. The secondary structure prediction certainly has room for improvement, although it is not clear how this could be done rapidly. Given the observation of an exonic ECD for the *KCNA1* site[80], it would be interesting to extend this method to identify exonic ECDs from a range of vertebrate species. This would involve a different method of randomisation than the one that was previously used. Finally, it could be very productive to compare the results of this screen with the results of the currently unpublished results of the Ohman groups ADAR2 antibody pull-down screen. As a completely independent source of data, this could provide valuable validation for my ECD predictions.

7.4 Conserved ECD Screen in the Fruit Fly

Based on the success of the previous screen in vertebrates, and the abundance of the known edited sites in *Drosophila*, a screen was performed between *D. melanogaster* and *D. pseudoobscura*. Although this work is currently unpublished, it would be the first bioinformatic analysis of ECDs in this genus. Only two of the 63 known edited regions had published and experimentally validated ECD predictions. Our protocol was able to identify both of these as the best ECDs for their respective exons, although the second site was only found after extending the screen beyond 2kb of intronic sequence. This demonstrated that this protocol was able to identify genuine ECDs in *Drosophila*. ECD predictions were also generated for the other known edited sites in *Drosophila*. Thresholds were set that resulted in specificities of 97.7% and 98.4% for intronic and exonic ECDs, respectively. Thirty-three of the known edited sites had ECDs below these thresholds that overlapped the edited sequences. This represents are sensitivity of 52%, although experimental data would be required to validate each of these ECDs. Some of the possible reasons why almost half of these edited sites did not have high quality overlapping ECD predictions are discussed at the end of Chapter 5. Interestingly, many of the known edited sites were overlapped by more than one ECD below these strict thresholds. It is possible that each of these putative duplexes can form and interfere with each other to regulate editing. Unfortunately, extensive experimental data would be required to demonstrate this.

In addition to analysing the known edited sites, this protocol was also used to screen these *Drosophila* genomes for novel edited sites. The criteria used in these screens were stricter than those used for the known sites alone. This resulted in a number of intronic and exonic candidate ECDs that might underlie edited sites. In agreement with the known edited sites, these candidates include many genes that encode channels and receptors that are involved in the nervous system. In particular there appears to be a group of genes involved in axon development and function. Some of these candidates are being experimentally tested for evidence of editing.

As with the previous chapter, there are several caveats to this method. These are discussed towards the end of Chapter 5. The most important caveat, however, is that there is a degree of circularity involved in this protocol, which could have resulted in inflated estimates of sensitivity. The reason for this is that our protocol relies heavily on sequence conservation and many of the known edited sites were identified through the observation of very high sequence conservation[140]. It is not clear how this could have been avoided, as high levels of sequence conservation seem integral to most of the *Drosophila* sites.

7.5 A Comparison of Vertebrate and *Drosophila* Recoding Edited Sites

Table 7.1 (p164) shows some of the similarities and differences between the known vertebrate and *Drosophila* recoding edited sites. Broadly these two sets of sites are very similar. There are noticeable differences however. There are many more *Drosophila* sites, and a much higher proportion of the edited genes contain more than one edited exon. The majority of the vertebrate edited sites occur at the 3'end of an exon. In contrast, the *Drosophila* sites are predominantly in the middle of the exon. The ECS locations also differ with most of the vertebrate ECSs in the 3'intron, and the *Drosophila* ECS predictions being spread almost equally between the exon, the 5'intron and the 3'intron. Although these differences are quite clear, the reasons behind them are not apparent.

Table 7.1. A Comparison of the Known Recoding Edited Sites in Vertebrates and *Drosophila*.

Feature	The Known Vertebrate Recoding Sites	The Known *Drosophila* Recoding Sites
Quantity	16 sites/clusters	63 sites/clusters
Gene Functions	Most sites are in channels or receptors involved in neuronal function[1].	Most sites are involved in the adult nervous system. Especially genes involved in rapid electrical and chemical neurotransmission[34,140].
Sequence Conservation of ESs and ECSs	High (81%-99% nucleotide identity between mouse & human)	Very high (typically >98% of 50bp between *D.mel* and *D.psu*)[140]
Clustering of Sites	Yes (6/16).	Yes (22/63).
Multiple Edited Exons per Gene	Yes (6/16).	Yes (50/63).
Multiple ECD predictions per Site	Multiple ECD predictions not observed in vertebrate screen.	Yes (approx. 14/63)
ECD Lengths	Varies between ~29bp (*GluR-5* site) & ~120bp (*BC10*)	Currently unknown. Many appear > 20bp.
Quality of ECD Secondary Structure	Imperfect	Imperfect (and of similar quality)
ES Locations	9/12 experimentally validated ECDs overlap 3' end of exon.	Of 37 sites with good ECDs, the edited sites are >25bp from either end in 22 sites. 8 sites are near the 5'end, 7 sites are near the 3'end.
ECS Locations	1/12 experimentally validated ECDs in exon, 2 in 5'intron, 9 in 3'intron.	Of 37 sites with good ECDs, there are 16 exonic ECSs, 12 5prime intronic ECSs & 13 3prime intronic ECSs. (Some sites have >1 ECD).

7.6 Editing and Splicing

A number of arguments were provided in the Introduction that suggest there is a link between A-I RNA editing and splicing. Several aspects of the data in this thesis also support this link. For example, the observation of a highly conserved ECD overlapping the branch site for all four AMPA receptor subunit flip exons suggests that editing of this site might interfere with splicing. If this duplex does not regulate splicing in some way, it is unclear what purpose it may have, as it does not overlap the coding sequence in the exon. Many of the novel candidates in the vertebrate screen appear to be involved in splicing and one gene involved in splicing, *Kuzbanian*, was found in both the intronic and exonic *Drosophila* ECD screens. Additionally, there were many novel vertebrate ECDs that overlapped the exon boundaries, although this number is not statistically significant. Some of the ECD predictions covered both ends of the exons they were found in. The splicing mechanism recently described for *DsCam*[120] demonstrates that editing and splicing both use conserved duplexes, which raises the possibility that each of these ECD predictions may be involved in editing, splicing, both or neither.

7.7 The Evolution of Editing

It was hoped that these analyses might help to explain the evolution of edited sites and their ECDs. However, it was only possible to make very simple observations on these processes. Generally, the ECD structures that I have been able to identify show very little change in either sequence or structure between species. Due to the nature of this protocol, any ECDs that had changed significantly would not have been detected. This is one possible explanation for why ECDs were not predicted for all of the known edited sites. Notably, however, where changes do occur, there is often a compensatory change in the opposite base. An example of this can be seen for BC10 in Figure 3.4 (p66 - first and last mismatches in top half of alignment).

7.8 Future Directions

Although the sensitivities and specificities of these analyses were all very respectable, there is always room for improvement. There are a number of existing

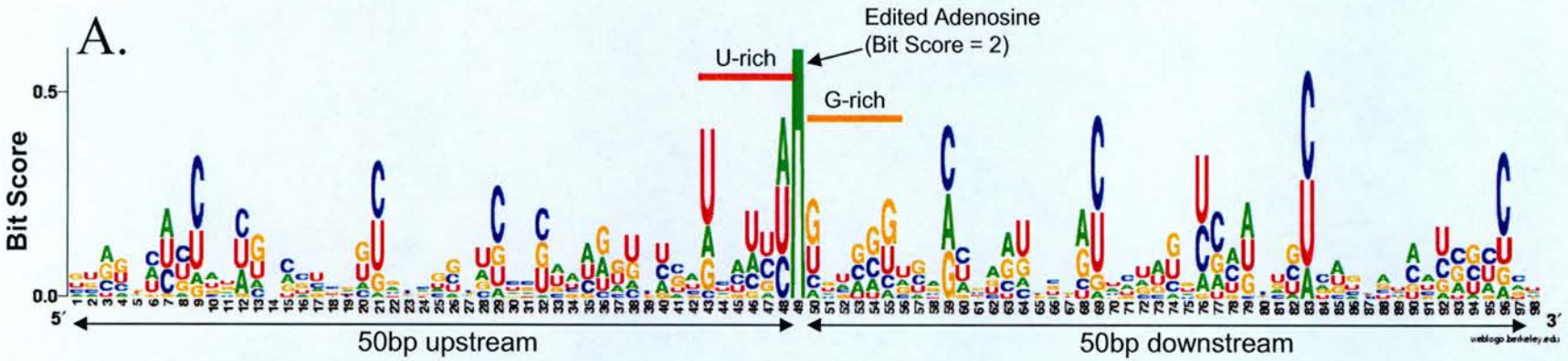
features that could be improved, or new features that could be added to improve these analyses. The method for finding ECDs could probably be improved, although it is not clear how this would be achieved. A method that compromised between the high speed of this approach, and the accuracy of RNA folding algorithms could be more suitable, especially as computer processor speeds are continually getting faster. The information that ensures the ADARs only edit the correct adenosines in the correct transcript must exist, otherwise every RNA duplex would be edited. Identifying and accurately characterising the sources of this specificity would be extremely useful for predicting any remaining unknown sites.

The ECD prediction method used in the mismatch-based screen was relatively very simple. Using an enhanced ECD prediction method could result in greatly improved predictions. It would also be possible to include additional features to those described in Chapter 3. One possibility would be to use the base preferences of the nucleotides on either side of the known edited sites. For example, the preceding base is most often a U or an A in ADAR2 edited sites (according to Dawson *et al*)[63]. Another possibility would be to use the predicted ECD structure to determine what base is opposite the putatively edited base. It has been shown that there is a preference for the opposite base to be a C[92].

To investigate the possibility that there are other features of the edited sequences I aligned all the protein recoding A-I edited sequences and looked for any useful patterns. Figure 7.1 (p167) shows the relative abundance of each nucleotide in the 50bp flanking either side of the edited adenosine. It appears that there are many positions that are enriched for cytosine and uracils. These positions are well spread out, almost appearing periodically throughout the sequence. It is possible that this represents a preference for these bases on one side of a helical RNA duplex structure. These observations are preliminary, but could underlie an interesting and useful result.

It is also possible that there are novel motifs or structures elsewhere in the transcripts or promoters of these genes. It would be possible to scan these sequences for such a motif using MEME[188], or a similar motif detection algorithm. Based on the observation that ADAR1 can be found in the cytoplasm[181], it is also possible that some ECDs could occur in mature mRNAs, although there is currently no evidence for this. This is another possible explanation for the absence of ECD predictions

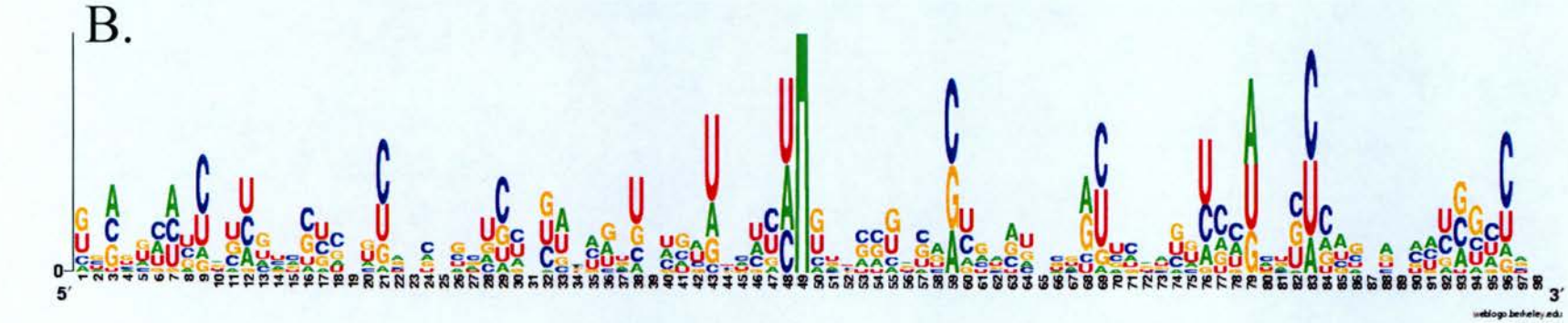
Figure 7.1. Sequence Composition Around the Known Mammalian Protein Recoding Edited Sites



$$2 + \sum_{b=\{A,C,G,T\}} p(b)\log_2 p(b)$$

Equation for calculating the bit score.

Section A shows the sequence composition in the 100bp around the edited adenosine. The large A in the middle of the figure is the edited adenosine. Only sites that result in known protein recoding edits are included in this analysis. There appears to be a short U-rich region before the edited adenosine and a short G-rich region immediately after it. Additionally, there are many positions that are enriched for C's or U's. These positions appear to be well spread out and may even be periodical. In order to ensure that these features are not a result of many similar sequences, we removed all but one of the GluR R/G and Q/R sites. Section B shows the results of this analysis. Both figures show the same periodically conserved C's or U's, but the second analysis does not have the U-rich and G-rich regions around the edited nucleotide. These figures were generated using WebLogo (Crooks *et al.*, *Genome Research*, 14:1188-1190).



from some of the known edited sites as the ECS would not be located in an adjacent intron.

In addition to improving these analyses, it would also be interesting to apply them to different pairs and groups of species. Two obvious comparisons would be to compare pufferfish with zebrafish, and *Caenorhabditis eleg*ans with *Caenorhabditis briggsae.* Editing sites have been observed in the nematode[2,189] and both fish[190], and all four species have genome assemblies and gene annotation publicly available. In theory these analyses could also be applied to plants and other more distant species, although it is not clear how successful they would be. In my absence these analyses could be carried out by anyone, using the application described in Chapter 6.

I have also considered applying these methods to identify C-U RNA edited sites. Mammalian C-U editing has only been shown to occur in mature mRNAs and only two targets are known (ApoB & NF1)[1]. This small number of known sites would make analysis of the success of this method very difficult.

7.9 Summary

I have successfully identified and validated *BC10* as a novel protein recoding edited gene in mammals[79]. Evidence has also been provided to support novel candidate edited sites from both vertebrates and *Drosophila.* I have also added to the characterisation of the known edited sites in both vertebrates and *Drosophila*, by analysing their sequence conservation and their ECDs (where they could be identified). This work has also added to our understanding of editing sites and their ECDs, including the evolution of editing, the specificity of editing and the link between editing and splicing.

Appendix 1. Additional Vertebrate ECD Figures

This appendix contains figures for the 30 top-scoring vertebrate ECD predictions. Each figure provides the Ensembl gene name, gene description, gene function, ECD LOD score, ECD rank, and a multi-species alignment for the ECD. The species that contain the illustrated ECDs are indicated in the 'Species Depth' section. Where the ECD is palindromic this is indicated below the alignment. The location of the ECS is described in the next section. The possible locations are 'Intronic', 'Exonic' and 'Boundary'. These are described in Chapter 4. The distance between the ES and the ECS sequences is also included ('ECD Half Separation'). The 'Additional Information' section describes the related known and candidate ECDs and whether there are any A-G mismatches to support the ECD. These figures are also available on the attached CD.

Figure A1.1. Candidate Report for ENSMUSE00000478281

ENSEMBL GENE: **ENSMUSG00000001986**
DESCRIPTION: **Glutamate receptor 3 precursor (GluR-3/GluR-C/AMPA3)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **1st** *LOD SCORE*: **35.8**
SPECIES DEPTH: [M] [R] [H] [G] [D] [F]

CORE ECD ALIGNMENT

```
Fugu R.    GCAUUAAGGUGGGUGGAAUAAUAUAACAAUAUUCUCCAUGUUGUUAUAGUAUUCCACC
Danio R.   GCAUUAAGGUGGGUGGAAUAAUAUAACAAUAUUGUCCAUGUUGUUAUAGUAUUCCACC
Gallus G.  GCAUUAAGGUGGGUGGAAUAAUAUAACAAUAUCC---AUGUUGUUAUAGUAUUCCACC
Homo S.    GCAUUAAGGUGGGUGGAAUAAUAUAACAAUAUCC---GUGUUGUUAUAGUAUUCCACC
Rattus N.  GCAUUAAGGUGGGUGGAAUAAUAUAACAAUAUCC---GUGUUGUUAUAAUAUUCCACC
Mus M.     GCAUUAAGGUGGGUGGAAUAAUAUAACAAUAUCC---GUGUUGUUAUAGUAUUCCACC
           |||||| |||||||||||| |||||||||||     |||||||||| |||||||||
Mus M.     CGUAGUCCCAUCCACCUUAUGAUAUUGUUGUGCC---UAUAACAAUAUAAUAAGGUGG
Rattus N.  CGUAGUCCCAUCCACCUUAUAAUAUUGUUGUGCC---UAUAACAAUAUAAUAAGGUGG
Homo S.    CGUAGUCCCAUCCACCUUAUGAUAUUGUUGUGCC---UAUAACAAUAUAAUAAGGUGG
Gallus G.  UGUAGUCCCAUCCACCUUAUGAUAUUGUUGUACC---UAUAACAAUAUAAUAAGGUGG
Danio R.   UGUAGUCCCAUCCACCUUAUGAUAUUGUUGUACCUGUUAUAACAAUAUAAUAAGGUGG
Fugu R.    UGUAGUCCCAUCCACCUUAUGAUAUUGUUGUA-----UAUAACAAUAUAAUAAGGUGG
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Boundary**: ECD Half Separation=**4bp**: Where=**3 prime/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (4th) GluR-B R/G Site
(6th) GluR-D R/G Site
(3rd) GluR-C Novel Site

A/G MISMATCHES: No A-G Mismatches observed.

Figure A1.2 Candidate Report for ENSMUSE00000477286

ENSEMBL GENE: **ENSMUSG00000033981**
DESCRIPTION: **Glutamate receptor 2 precursor (GluR-2/GluR-B/GluR-K2/AMPA2)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **2nd** *LOD SCORE*: **34.6**
SPECIES DEPTH: [M] [R] [H] [G] [X] [F]

CORE ECD ALIGNMENT

```
Fugu R.    AAAAUAACAUUGAUAAUGUUAUUU-AUGUUAUUUU
Gallus G.  AAAAUAACAUUGAUAAUGUUAUUU-AUGUUAUUUU
Homo S.    AAAAUAACAUUGAUAAUGUUAUUU-AUGUUAUUUU
Rattus N.  AAAAUAACAUUGAUAAUGUUAUUU-AUGUUAUUUU
Mus M.     AAAAUAACAUUGAUAAUGUUAUUU-AUGUUAUUUU
           |||||||||| ||||||||||||  ||||||||||
Mus M.     UUUUAUUGUAUUUAUUGUAAUAGU-UACAAUAAAA
Rattus N.  UUUUAUUGUAUUUAUUGUAAUAGU-UACAAUAAAA
Homo S.    UUUUAUUGUAUUUAUUGUAAUAGU-UACAAUAAAA
Gallus G.  UUU-AUUGUAAUAGUUACAAUAAAAUACAAUAAGU
Fugu R.    UUU-AUUGUAAUAGUUACAAUAAAAUACAAUAAUA
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Intronic**: ECD Half Separation=**0bp**: Where=**5 prime/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: (4th) GluR-B R/G Site
(5th) GluR-B Q/R Site
(7th) GluR-A Novel Site
(3rd) GluR-C Novel Site
(8th) GluR-D Novel Site

A/G MISMATCHES: No A-G Mismatches observed.

Figure A1.3 Candidate Report for ENSMUSE00000490396

ENSEMBL GENE: **ENSMUSG00000001986**
DESCRIPTION: **Glutamate receptor 3 precursor (GluR-3/GluR-C/AMPA3)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **3rd** *LOD SCORE*: **34.6**
SPECIES DEPTH: [M] [R] [H] [G] [D] [F]

CORE ECD ALIGNMENT

```
Fugu R.     AAAUAACAUAAUAUAAUGUUAU-UUAUGUUAUUU
Danio R.    AAAUAACAUAAUAUAAUGUUAU-UUAUGUUAUUU
Gallus G.   AAAUAACAUAAUAUAAUGUUAU-UUAUGUUAUUU
Homo S.     AAAUAACAUAAUAUAAUGUUAU-UUAUGUUAUUU
Rattus N.   AAAUAACAUAAUAUAAUGUUAU-UUAUGUUAUUU
Mus M.      AAAUAACAUAAUAUAAUGUUAU-UUAUGUUAUUU
            ||||||||||| |||||||||| ||||||||||
Mus M.      UUUAUUGUAUU-UAUUGUAAUAUAAUACAAUAAA
Rattus N.   UUUAUUGUAUU-UAUUGUAAUAUAAUACAAUAAA
Homo S.     UUUAUUGUAUU-UAUUGUAAUAUAAUACAAUAAA
Gallus G.   UUUAUUGUAUU-UAUUGUAAUAUAAUACAAUAAA
Danio R.    UUUAUUGUAUU-UAUUGUAAUAUAAUACAAUAAA
Fugu R.     UUUAUUGUAUU-UAUUGUAAUAUAAUACAAUAAA
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Intronic**: ECD Half Separation=**0bp**: Where=**5 prime/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: (1st) GluR-C R/G Site
(7th) GluR-A Novel Site
(2nd) GluR-B Novel Site
(8th) GluR-D Novel Site

A/G MISMATCHES: No A-G Mismatches observed.

Figure A1.4 Candidate Report for ENSMUSE00000223221

ENSEMBL GENE: **ENSMUSG00000033981**
DESCRIPTION: **Glutamate receptor 2 precursor (GluR-2/GluR-B/GluR-K2/AMPA2)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **4th** *LOD SCORE*: **32.7**
SPECIES DEPTH: [M] [R] [H] [G] [D] [F]

CORE ECD ALIGNMENT

```
Fugu R.     CAUUAAGGUGGGUGGAAUAGUAUAACAAUGUGUCCAAUGUUGUUAUAGUAUCCCACC
Danio R.    CAUUAAGGUGGGUGGAAUAGUAUAACAAUGUGUCCAAUGUUGUUAUAGUAUCCCACC
Gallus G.   CAUUAAGGUGGGUGGAAUAGUAUAACAAUAUGCUAAAUGUUGUUAUAGUAUCCCACC
Homo S.     CAUUAAGGUGGGUGGAAUAGUAUAACAAUAUGCUAAAUGUUGUUAUAGUAUCCCACC
Rattus N.   CAUUAAGGUGGGUGGAAUAGUAUAACAAUGUGCUCAAUGUUGUUAUAGUAUCCCACC
Mus M.      CAUUAAGGUGGGUGGAAUAGUAUAACAAUAUGCUCAAUGUUGUUAUAGUAUCCCACC
            ||||| ||||||||| |||||||||||||||     |||||||||| ||| |||||
Mus M.      GUAGUCCCAUCCACCCUAUGAUAUUGUUGUAACUCGUAUAACAAUAUGAUAAGGUGG
Rattus N.   GUAGUCCCAUCCACCCUAUGAUAUUGUUGUAACUCGUGUAACAAUAUGAUAAGGUGG
Homo S.     GUAGUCCCAUCCACCCUAUGAUAUUGUUGUAAAUCGUAUAACAAUAUGAUAAGGUGG
Gallus G.   GUAGUCCCAUCCACCCUAUGAUAUUGUUGUAAAUCGUAUAACAAUAUGAUAAGGUGG
Danio R.    GUAGUCCCAUCCACCCUAUGAUAUUGUUGUAACCUGUGUAACAAUAUGAUAAGGUGG
Fugu R.     GUAGUCCCAUCCACCCUAUGAUAUUGUUGUAACCUGUGUAACAAUAUGAUAAGGUGG
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Boundary**: ECD Half Separation=**17bp**: Where=**3 prime/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (1st) GluR-C R/G Site
(6th) GluR-D R/G Site
(5th) GluR-B Q/R Site
(2nd) GluR-B Novel Site

A/G MISMATCHES: No A-G Mismatches observed.

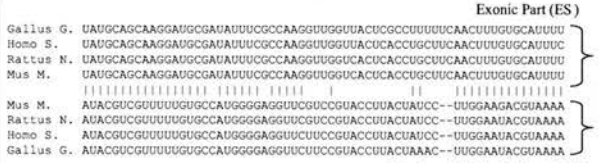
Figure A1.5 Candidate Report for ENSMUSE00000223233

ENSEMBL GENE: **ENSMUSG00000033981**
DESCRIPTION: **Glutamate receptor 2 precursor (GluR-2/GluR-B/GluR-K2/AMPA2)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **5th** *LOD SCORE*: **25.5**
SPECIES DEPTH: M R H G ☒ ☒

CORE ECD ALIGNMENT



ECS LOCATION Type=**Standard**: ECD Half Separation=**280bp**: Where=**Exon/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (42nd) GluR-6 Q/R Site
(4th) GluR-B R/G Site
(2nd) GluR-B Novel Site

A/G MISMATCHES: A-G Mismatches overlap the ES.

Figure A1.6 Candidate Report for ENSMUSE00000519849

ENSEMBL GENE: **ENSMUSG00000025892**
DESCRIPTION: **Glutamate receptor 4 precursor (GluR-4)/GluR4/GluR-D/AMPA4)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **6th** *LOD SCORE*: **24.4**
SPECIES DEPTH: M ☒ H G D F

CORE ECD ALIGNMENT

ECS LOCATION Type=**Boundary**: ECD Half Separation=**11bp**: Where=**3 prime/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (4th) GluR-B R/G Site
(1st) GluR-C R/G Site
(8th) GluR-D Novel Site

A/G MISMATCHES: No A-G Mismatches observed.

Figure A1.7 Candidate Report for ENSMUSE00000105648

ENSEMBL GENE: **ENSMUSG00000020524**
DESCRIPTION: **Glutamate receptor 1 precursor (GluR-1/GluR-A/GluR-K1/AMPA1)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **7th** *LOD SCORE*: **20.5**
SPECIES DEPTH: [M] [H] [G]

CORE ECD ALIGNMENT

```
Gallus G. AAAUAACAUUGGUAAUGUUAUUU-AUGUUAUUU
Homo S.   AAAUAACAUUGGUAAUGUUAUUU-AUGUUAUUU
Mus M.    AAAUAACAUUGGUAAUGUUAUUU-AUGUUAUUU
          ||||||||| |||||||||||  |||||||||
Mus M.    UUUAUUGUAUUUAUUGUAAUGGU-UACAAUAAA
Homo S.   UUUAUUGUAUUUAUUGUAAUGGU-UACAAUAAA
Gallus G. UU-AUUGUAAUGGUUACAAUAAAAUACAAUAAG
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Intronic**: ECD Half Separation=**0bp**: Where=**5 prime/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: (2nd) GluR-B Novel Site
(3rd) GluR-C Novel Site
(8th) GluR-C Novel Site

A/G MISMATCHES: No A-G Mismatches observed.

Figure A1.8 Candidate Report for ENSMUSE00000515006

ENSEMBL GENE: **ENSMUSG00000025892**
DESCRIPTION: **Glutamate receptor 4 precursor (GluR-4)/GluR4/GluR-D/AMPA4)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **8th** *LOD SCORE*: **20.5**
SPECIES DEPTH: [M] [H] [G]

CORE ECD ALIGNMENT

```
Gallus G. AAAAUAACAUAUAUAAUGUUAUUUAUGUUAUUUU
Homo S.   AAAAUAACAUAUAUAAUGUUAUUUAUGUUAUUUU
Mus M.    AAAAUAACAUAUAUAAUGUUAUUUAUGUUAUUUU
          |||||||||| ||||||||| ||||||||||
Mus M.    UUUUAUUGUAUUUAUUGUAAUAUAUACAAUAAAA
Homo S.   UUUUAUUGUAUUUAUUGUAAUAUAUACAAUAAAA
Gallus G. UUUUAUUGUAUUUAUUGUAAUAUAUACAAUAAAA
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Intronic**: ECD Half Separation=**0bp**: Where=**5 prime/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: (6th) GluR-D R/G Site
(7th) GluR-A Novel Site
(2nd) GluR-B Novel Site
(3rd) GluR-C Novel Site

A/G MISMATCHES: No A-G Mismatches observed.

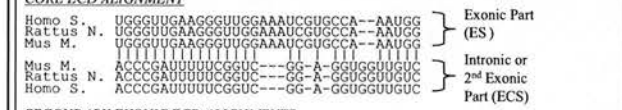
Figure A1.9 Candidate Report for ENSMUSE00000479443

ENSEMBL GENE: **ENSMUSG00000007850**
DESCRIPTION: **Heterogeneous nuclear ribonucleoprotein H (hnRNP H)**

FUNCTION: Involved in pre-mRNA processing.

RANK: **9th** *LOD SCORE*: **20.1**
SPECIES DEPTH: [M] [R] [H] [X] [X] [X] & [M] [R] [H] [G] [D] [X]

CORE ECD ALIGNMENT



Exonic Part (ES)

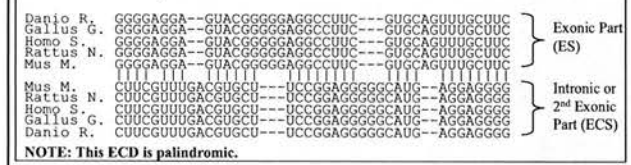
Intronic or 2nd Exonic Part (ECS)

SECONDARY EXONIC ECD ALIGNMENTS

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.



Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Boundary**: ECD Half Separation=**0bp**: Where=**5 prime/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: A-G mismatch observed nearby ES.

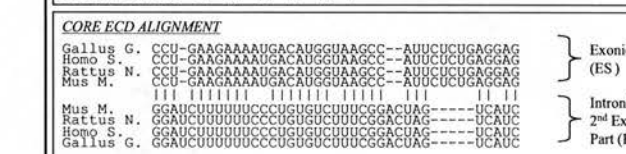
Figure A1.10 Candidate Report for ENSMUSE00000113629

ENSEMBL GENE: **ENSMUSG00000047022**
DESCRIPTION: **Mirror-image polydactyly gene 1 homolog**

FUNCTION: Anomalies in this gene can result in mirror-image duplication of digits.

RANK: **10th** *LOD SCORE*: **20.0**
SPECIES DEPTH: [M] [R] [H] [G] [X] [X]

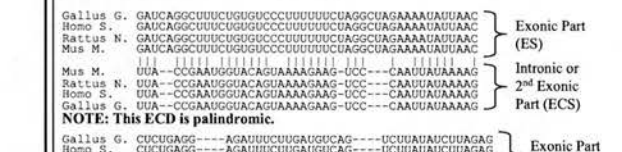
CORE ECD ALIGNMENT



Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

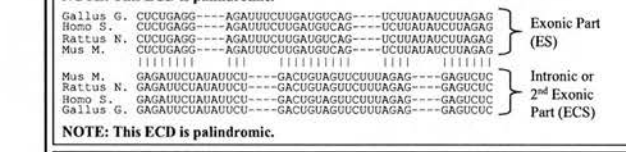
SECONDARY ECD ALIGNMENTS (EXONIC)



Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.



Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Boundary**: ECD Half Separation=**16bp**: Where=**3 prime/5 prime**
The ECD halves flank the exon closely, overlapping both exon junctions.

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G Mismatches observed.

Figure A1.11 Candidate Report for ENSMUSE00000228059

ENSEMBL GENE: **ENSMUSG00000033569**
DESCRIPTION: **Brain-specific angiogenesis inhibitor 3 precursor**

FUNCTION: G-protein coupled receptor involved in brain-specific angiogenesis regulation.

RANK: **11th** *LOD SCORE*: **20.0**
SPECIES DEPTH: M R H G ☒ ☒

CORE ECD ALIGNMENT

```
Gallus G.  AGUAGGUGC-AAAGCCAGCUUUAGUACUUGCU
Homo S.    AGUAGGUGC-AAAGCCAGCUUUAGUACUUGCU
Rattus N.  AGUAGGUGC-AAAGCCAGCUUUAGUACUUGCU
Mus M.     AGUAGGUGC-AAAGCCAGCUUUAGUACUUGCU
           ||||||||| ||||| ||||| |||||||||
Mus M.     UCGUUCAUGAUUUCGACCGAAA-CGUGGAUGA
Rattus N.  UCGUUCAUGAUUUCGACCGAAA-CGUGGAUGA
Homo S.    UCGUUCAUGAUUUCGACCGAAA-CGUGGAUGA
Gallus G.  UCGUUCAUGAUUUCGACCGAAA-CGUGGAUGA
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

ECS LOCATION Type=**Boundary**: ECD Half Separation=**0bp**: Where=**3 prime/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.12 Candidate Report for ENSMUSE00000177912

ENSEMBL GENE: **ENSMUSG00000028289**
DESCRIPTION: **Ephrin type-A receptor 7 precursor (Tyrosine-protein kinase receptor EHK-3)**
FUNCTION: Ephrin receptor with tyrosine kinase activity.

RANK: **12th** *LOD SCORE*: **20.0**
SPECIES DEPTH: M R H G ☒ ☒

CORE ECD ALIGNMENT

```
Gallus G.  GAGCGUGUGAUUG--GUGCAGGUAAGGC---UGUCGU
Homo S.    GAGCGUGUGAUUG--GUGCAGGUAAGGC---UAUUGU
Rattus N.  GAGCGUGUGAUUG--GUGCAGGUAAGGC---UGUUGU
Mus M.     GAGCGUGUGAUUG--GUGCAGGUAAGGC---UGUUGU
           ||| ||||||||| ||| |||||||||   ||||||
Mus M.     CUUUCGUACUGACACUAC-UUCGUUUCGUGAAUAACA
Rattus N.  CUUUCGUACUGACACUAC-UUCGUUUCGUGAAUAACA
Homo S.    CUUUCGUACUGACACUAC-UUCGUUUCGUGAAUAACA
Gallus G.  CUUUCGUACUGACACUAC-UUCGUUUC--------CG
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

SECONDARY ECD ALIGNMENT (EXONIC)

```
Gallus G.  --AGUCAUGCUUU-CUUACAUCA-UUACA-GUUAAAUUU
Homo S.    ACAGUCAUGCUUU-CUUACAUCA-UUACA-GUUAAAUUU
Rattus N.  ACAGUCAUGCUUU-CUUACAUCA-UUACA-GUUAAAUUU
Mus M.     ACAGUCAUGCUUU-CUUACAUCA-UUACA-GUUAAAUUU
           |||| || ||||| ||| ||||| ||||| |||||||||
Mus M.     UGUU-GU-CGGAAUGGACGUGGUUAGUGUGCGAGUUAAA
Rattus N.  UGUU-GU-CGGAAUGGACGUGGUUAGUGUGCGAGUUAAA
Homo S.    UGUU-AU-CGGAAUGGACGUGGUUAGUGUGCGAGUUAAA
Gallus G.  GUGCUGU-CGGAAUGGACGUGGUUAGUGUGCGAGUUAAA
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Boundary**: ECD Half Separation=**122bp**: Where=**3 prime/5 prime**

The ECD halves flank the exon closely, overlapping both exon junctions.

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.13 Candidate Report for ENSMUSE00000113867

ENSEMBL GENE: **ENSMUSG00000061603**
DESCRIPTION: **NA**

FUNCTION: A kinase (PRKA) anchor protein 6. Binds to type II regulatory subunits of protein kinase A and anchors/targets them to the nuclear membrane or sarcoplasmic reticulum.

RANK: **13th** *LOD SCORE*: **20.0**
SPECIES DEPTH: [M] [X] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G. GUAUUUGCAUUUUUAUUACUGUUUG---UGGGU  } Exonic Part
Homo S.   GUAUUUGCAUUUUUAUUACUGUUUG---UAGGU  } (ES)
Mus M.    GUAUUUGCAUUUUUAUUACUGUUUG---UAGGU
          |||||| |||||| || ||||||||   |||||
Mus M.    CAUAAAGGUAAGA-UG-UGACAAGUGGAAUCUA  } Intronic or
Homo S.   CAUAAAGGUAAGG-UG-UGACAAGUGGAAUCUA  } 2nd Exonic
Gallus G. CAUAAAGGUAAGG-UG-UGACAAGUGGAAUCUA  } Part (ECS)
```

ECS LOCATION Type=**Boundary**: ECD Half Separation=**2,152bp**: Where=**3 prime/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.14 Candidate Report for ENSMUSE00000186519

ENSEMBL GENE: **ENSMUSG00000029169**
DESCRIPTION: **Putative pre-mRNA splicing factor RNA helicase (DEAH box protein 15)**

FUNCTION: Pre-mRNA processing factor involved in disassembly of spliceosomes after the release of mature mRNA

RANK: **14th** *LOD SCORE*: **20.0**
SPECIES DEPTH: [M] [R] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G. AUGGGCGUUUGUGUUGGGCCCA-UACCAAC-CUUGU  } Exonic Part
Homo S.   AUGGGCGUUUGUGUUGGGCCCA-UACCAAC-CUUGU  } (ES)
Rattus N. AUGGGCGUUUGUGUUGGGCCCA-UACCAAC-CUUGU  }
Mus M.    AUGGGCGUUUGUGUUGGGCCCA-UACCAAC-CUUGU  }
          ||||| ||| |||| ||||||| ||| ||| |||||
Mus M.    UGUUC-CAACCAUA-CCCGGGUUGUGUUUGCGGGUA  } Intronic or
Rattus N. UGUUC-CAACCAUA-CCCGGGUUGUGUUUGCGGGUA  } 2nd Exonic
Homo S.   UGUUC-CAACCAUA-CCCGGGUUGUGUUUGCGGGUA  } Part (ECS)
Gallus G. UGUUC-CAACCAUA-CCCGGGUUGUGUUUGCGGGUA  }
```

ECS LOCATION Type=**Standard**: ECD Half Separation=**98bp**: Where=**Exon/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.15 Candidate Report for ENSMUSE00000496440

ENSEMBL GENE: **ENSMUSG00000029563**
DESCRIPTION: **Forkhead box protein P2**
FUNCTION: Transcriptional repressor that may play a role in the specification and differentiation of lung epithelium. May also play a role in developing neural, gastrointestinal and cardiovascular tissues. Involved in neural mechanisms mediating the development of speech and language.

RANK: **15th** *LOD SCORE*: **20.0**
SPECIES DEPTH: M R H G ☒ ☒

CORE ECD ALIGNMENT

```
Gallus G. GUUUUGUUCGAG-UAG-AAAAUGUUAAAGGAGCAGUAU
Homo S.   GUUUUGUUCGAG-UAG-AAAAUGUUAAAGGAGCAGUAU
Rattus N. GUUUUGUUCGAG-UAG-AAAAUGUUAAAGGAGCAGUGU
Mus M.    GUUUUGUUCGAG-UAG-AAAAUGUUAAAGGAGCAGUGU
          |||||||| ||| ||  ||||||||||||| || |||
Mus M.    CAGGACGAACUCCAUAAUUUUACGAUUUUUCCGAUGUG
Rattus N. CAGGACGAACUCCAUAAUUUUACGAUUUUUCCGAUGUC
Homo S.   CAGGACGAACUCCAUAAUUUUACGAUUUUUCCGAUGUC
Gallus G. CAGGACGAACUCCAUAAUUUUACGAUUUUUCCGAUGUC
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Standard**: ECD Half Separation=**412bp**: Where=**Exon/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.16 Candidate Report for ENSMUSE00000375270

ENSEMBL GENE: **ENSMUSG00000028546**
DESCRIPTION: **ELAV-like protein 4 (Paraneoplastic encephalomyelitis antigen HuD)**

FUNCTION: Belongs to the RNP elav family. May play a role in neuron-specific RNA processing. Drosophila Elav-like 4 (embryonic,lethal,abnormal vision),neural specific,expressed in brain,identified using sera from patients with a paraplastic multifocal neurologic disorder,homologous to HD,shuttling the ERG class of mRNA to the cytoplasm in response to regulatory signals,where they become stabilized,translated or rapidly degraded

RANK: **16th** *LOD SCORE*: **19.3**
SPECIES DEPTH: M R H ☒ ☒ ☒

CORE ECD ALIGNMENT

```
Homo S.   GAGGUGUACCAGGAAAGGAU-UUUAUAAUGCUU
Rattus N. GAGGUGUACCAGGAAAGGAU-UUUAUAAUGCUU
Mus M.    GAGGUGUACCAGGAAAGGAU-UUUAUAAUGCUU
          ||||| ||  |||||||||| ||||||||||||
Mus M.    UUCUA-AUACUUUUUUUUUGUAAAUGUUAUGAA
Rattus N. UUCUA-AUACUUUUUUUUUGUAAAUGUUAUGAA
Homo S.   UUCUA-AUACUUUUUUUUUGUAAAUGUUAUGAA
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

SECONDARY ECS ALIGNMENT

```
Homo S.   GACAGAGU-GUUGCAAGUUUCCUUUAAA
Rattus N. GACAGAGU-GUUGCAAGUUUCCUUUAAA
Mus M.    GACAGAGU-GUUGCAAGUUUCCUUUAAA
          |||||||  |||   ||||||||||||
Mus M.    CUGUUUCAACAG---UUAAAGGGGAUUU
Rattus N. CUGUUUCAACAG---UUAAAGGGGAUUU
Homo S.   CUGUUUCAACAG---UUAAAGGGGAUUU
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Standard**: ECD Half Separation=**2,309bp**: Where=**Exon/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (34th/35th) Novel ECD in same gene.

A/G MISMATCHES: A-G mismatch observed nearby ES.

Figure A1.17 Candidate Report for ENSMUSE00000466378

ENSEMBL GENE: **ENSMUSG00000025255**
DESCRIPTION: **Zinc finger homeodomain 4**

FUNCTION: Homeobox transcription factor. Candidate gene for congenital bilateral isolated ptosis disorder.

RANK: **17th** *LOD SCORE*: **19.2**
SPECIES DEPTH: [M] [X] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G. AAAGUUGUUCAGAUGAGUCUGACAGUGAGUUGAGCCAGAAG
Homo S.   AAAGUUGUUCAGAUGAGUCUGACAGUGAGCUGAGCCAGAAG
Mus M.    AAAGUUGUUCAGAUGAGUCUGACAGUGAGCUGAGCCAGAAG
          ||||||||| || || ||||||| ||||||||| |||||
Mus M.    UUUCGAUGACUUAAUGUGGAUUG--AUUUGAUUC--UCUUC
Homo S.   UUUCGAUGACUUAAUGUGGAUUG--AUUUGAUUC--UCUUU
Gallus G. UUUCGAUGACUUAAUGUGGAUUG--AUUUGAUUC--UCUUU

Gallus G. UUUAUAUGCAGAAAGUGCC-AGAUGAUGUUUACUGAUGAAGA
Homo S.   UUUAUAUGCAGAAAGUGCC-AGAUGAUGUUUACUGAUGAAGA
Mus M.    UUUGUAUGCAGAAAGUGCC-AGAUGAUGUUUACUGACGAAGA
          ||||||||| ||| ||||| ||| ||||||| ||||||||||
Mus M.    AAACGUGUGAUUUGUGUGGAUCUUUUACGAAAGAUUAUUUCU
Homo S.   AAACAUGUGAUUUGUGUGGAUCUUUUACGAAAGAUUAUUUCU
Gallus G. AAACAUGUGAUUUGUGUGGAUCUUUUCCGAAAGAUUAUUUCU
```

Exonic Part (ES)
Intronic or 2nd Exonic Part (ECS)
Exonic Part (ES)
Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Standard**: ECD Half Separation=**501bp**: Where=**Exon/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (40^{th}) Novel ECS from same gene.
(18^{th}) Same exon.
A/G MISMATCHES: A-G mismatch observed nearby ES.

Figure A1.18 Candidate Report for ENSMUSE00000476281

ENSEMBL GENE: **ENSMUSG00000025255**
DESCRIPTION: **Zinc finger homeodomain 4**

FUNCTION: Homeobox transcription factor. Candidate gene for congenital bilateral isolated ptosis disorder.

RANK: **18th** *LOD SCORE*: **19.2**
SPECIES DEPTH: [M] [X] [H] [G] [X] [X]

Same exon as previous.

CORE ECD ALIGNMENT

```
Gallus G. AAAGUUGUUCAGAUGAGUCUGACAGUGACUUGAGCCAGAAG
Homo S.   AAAGUUGUUCAGAUGAGUCUGACAGUGACCUGAGCCAGAAG
Mus M.    AAAGUUGUUCAGAUGAGUCUGACAGUGACCUGAGCCAGAAG
          ||||||||| || || ||||||| ||||||||| |||||
Mus M.    UUUCGAUGACUUAAUGUGGAUUG--AUUCGAUUC--UCUUC
Homo S.   UUUCGAUGACUUAAUGUGGAUUG--AUUCGAUUC--UCUUU
Gallus G. UUUCGAUGACUUAAUGUGGAUUG--AUUCGAUUC--UCUUU

Gallus G. UUUAUAUGCAGAAAGUGCC-AGAUGAUGCUUACUGAUGAAGA
Homo S.   UUUAUAUGCAGAAAGUGCC-AGAUGAUGCUUACUGAUGAAGA
Mus M.    UUUGUAUGCAGAAAGUGCC-AGAUGAUGCUUACUGACGAAGA
          ||||||||| ||| ||||| ||| ||||||| ||||||||||
Mus M.    AAACGUGUGAUUUGUGUGGAUCUUUUACCAAAGAUUAUUUCU
Homo S.   AAACAUGUGAUUUGUGUGGAUCUUUUACCAAAGAUUAUUUCU
Gallus G. AAACAUGUGAUUUGUGUGGAUCUUUUCCCAAAGAUUAUUUCU
```

Exonic Part (ES)
Intronic or 2nd Exonic Part (ECS)
Exonic Part (ES)
Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Standard**: ECD Half Separation=**501bp**: Where=**Exon/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (40^{th}) Novel ECD from same gene.
(17^{th}) Same exon.
A/G MISMATCHES: A-G mismatch observed nearby ES.

Figure A1.19 Candidate Report for ENSMUSE00000472446

ENSEMBL GENE: **ENSMUSG00000057453**
DESCRIPTION: **Bladder cancer-associated protein (Bladder cancer 10 kDa protein) (Bc10)**

FUNCTION: Differential expression in bladder cancer. Unknown function.

RANK: **19th** *LOD SCORE*: **18.9**
SPECIES DEPTH: [M] [R] [H] [X] [X] [X]

CORE ECD ALIGNMENT

```
                                                                  Exonic Part (ES)
Homo S.    GCCCUGCCCGGCAGAGAUCAUGUAUUGCCUCCAGUGGCUGCUGCCCGUC  }
Rattus N.  GCCCUGCCCGGCAGAGAUCAUGUAUUGCCUCCAGUGGCUGCUGCCCGUC  }
Mus M.     GCCCUGCCCGGCAGAGAUCAUGUAUUGCCUCCAGUGGCUGCUGCCCGUC  }
           ||||||||||||||||||||||||||||||||| | |||||| |||||||
Mus M.     CGGGACGGGUCGUCUCUAGUACAUAGCGGAGGACCCCGACGUCGGGCAG  }
Rattus N.  CGGGACGGGUCGUCUCUAGUACAUAGCGUAGGACCCCGACGUCGGGCAG  }
Homo S.    CGGGACGGGUCGUCUCUAGUACAUAGCGGAGGACCCCGACGUCGGGCAG  }
                                                  Intronic or 2nd Exonic Part (ECS)
```

ECS LOCATION Type=**Standard**: ECD Half Separation=**450bp**: Where=**Exon/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: A-G mismatches overlapping the ES.

Figure A1.20 Candidate Report for ENSMUSE00000165486

ENSEMBL GENE: **ENSMUSG00000027018**
DESCRIPTION: **Histone/Histidine aminotransferase 1**

FUNCTION: May play a role in telomeric silencing. Acetylates soluble but not nucleosomal H4 and histone H2A.

RANK: **20th** *LOD SCORE*: **18.7**
SPECIES DEPTH: [M] [R] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G.  CUUGAUGUGGUUUAUUGAAACUGCUA-GCUUUAUUGA--UGUA-GAU  }
Homo S.    UUUGAUGUGGUUUAUUGAAACUGCUA-GCUUUAUUGACGUGGAUGAU  } Exonic Part
Rattus N.  UUUGAUGUGGUUUAUUGAAACUGCUA-GCUUUAUUGACGUGGAUGAU  } (ES)
Mus M.     UUUGAUGUGGUUUAUUGAAACUGCUA-GCUUUAUUGACGUGGAUGAU  }
           ||||||| ||| |||     |||||| | |||  ||| ||| |||||
Mus M.     GGGUUAC-CCAUGUG-----GACGAUAC-AAG--ACU-UACAUGUUA  }
Rattus N.  GGGUUAC-CCAUGUG-----GACGAUAC-AAG--ACU-UACAUGUUA  } Intronic or
Homo S.    GGGUUAC-CCAUGUG-----GACGAUAC-AAG--ACU-UACAUGUUA  } 2nd Exonic
Gallus G.  GGGUUAC-CCAUGUG-----GACGAUGC-AAG--ACU-UACAUGUUA  } Part (ECS)
```

ECS LOCATION Type=**Standard**: ECD Half Separation=**408bp**: Where=**Exon/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (41st) Novel ECS from same gene.

A/G MISMATCHES: A-G mismatch overlapping ES.

Figure A1.21 Candidate Report for ENSMUSE00000327165

ENSEMBL GENE: **ENSMUSG00000037369**
DESCRIPTION: **Ubiquitously transcribed X chromosome tetratricopeptide repeat protein**

FUNCTION: Ubiquitously transcribed Jumonji transcription factor, which escapes X inactivation. Contains tetratricopeptide repeats (TPR).

RANK: **21st** *LOD SCORE*: **18.7**
SPECIES DEPTH: M R H G ☒ ☒

CORE ECD ALIGNMENT

```
Gallus G.  GUGAGAAGUUGGGUUAUGUUCUGCAGGUGCCCAGCUU---UAAGGGU
Homo S.    GUGAGAAGUUGGGUUAUGUUCUGCAGGUGCCCAGCUU---UAAGGGU
Rattus N.  GUGAGAAGUUGGGUUAUGUUCUGCAGGUGCCCAGCUU---UAAGGGU
Mus M.     GUGAGAAGUUGGGUUAUGUUCUGCAAGUGCCCAGCUU---UAAGGGU
           ||||||||| ||||| || |  |||||   || ||||   |||||||
Mus M.     CAUUUUUCG-UUUAA-AUUG--UGUUC---GGAUGAACUCAUUCUCG
Rattus N.  CAUUUUUCG-UUUAA-AUUG--UGUUC---GGAUGAACUCAUUCUCG
Homo S.    CAUUUUUCG-UUUAA-AUUG--UGUUC---GGAUGAACUCAUUCUCG
Gallus G.  CAUUUUUCG-UUUAA-AUUG--UGUUC---GGAUGAAAUCAUUCUCG
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Boundary**: ECD Half Separation=**145bp**: Where=**3 prime/5 prime**
The ECD halves flank the exon closely, overlapping both exon junctions.

ADDITIONAL INFORMATION
RELATED SITES: None.
A/G MISMATCHES: No A-G mismatch observed.

Figure A1.22 Candidate Report for ENSMUSE00000195189

ENSEMBL GENE: **ENSMUSG00000030096**
DESCRIPTION: **Sodium and chloride-dependent taurine and beta-alanine transporter.**

FUNCTION: Na/Cl-dependent taurine transporter. Involved in neurotransmitter transport.

RANK: **22nd** *LOD SCORE*: **18.5**
SPECIES DEPTH: M R H G ☒ ☒

CORE ECD ALIGNMENT

```
Gallus G.  GGUGGCAGUGGGUGGUGGGCACUGCCGCC
Homo S.    GGUGGCAGUGGGUGGUGGGCACUGCCACC
Rattus N.  GGUGGCAGUGGGUGGUGGGCACUGCCACC
Mus M.     GGUGGCAGUGGGUGGUGGGCACUGCCACC
           |||||||||| | || | ||||||||||
Mus M.     CCACCGUCACGGGUGGUGGUGACGGUGG
Rattus N.  CCACCGUCACGGGUGGUGGUGACGGUGG
Homo S.    CCACCGUCACGGGUGGUGGUGACGGUGG
Gallus G.  CCGCCGUCACGGGUGGUGGUGACGGUGG
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

```
Gallus G.  UGUGGCAGAGUCAGGUACGA---UGGUAUUG--GU---GGCAGU
Homo S.    UGUGGCUGAGUCAGGUACGG---UGGUAUCG--GU---GGCAGU
Rattus N.  UGUGGCUGAGUCAGGUACGG---UGGUGUCG--GU---GGCAGU
Mus M.     UGUGGCUGAGUCAGGUACGG---UGGUGUCG--GU---GGCAGU
           |||||   ||||||||||||   ||||||||  ||   |||| |
Mus M.     ACGCU---UCGGUCUGUGUUUUGACCAUGGUGACAAGUCCGUAG
Rattus N.  ACGCU---UCGGUCUGUGUUUUGACCAUGGUGACAAGUCCGUAG
Homo S.    ACGCU---UCGGUCUGUGUUUUGACCAUGGUGACAAGUCCGUAG
Gallus G.  ACGCU---UCGGUCUGUGUUUUGACCAUGGUGACAAGUCCGUAG
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Intronic**: ECD Half Separation=**0bp**: Where=**3 prime/3 prime**

ADDITIONAL INFORMATION
RELATED SITES: None.
A/G MISMATCHES: No A-G mismatch observed.

Figure A1.23 Candidate Report for ENSMUSE00000207263

ENSEMBL GENE: **ENSMUSG00000006678**
DESCRIPTION: **DNA polymerase alpha catalytic subunit**

FUNCTION: Involved in DNA replication of nuclear genome

RANK: **23rd** *LOD SCORE*: **18.5**
SPECIES DEPTH: [M] [R] [H] [X] [X] [X]

CORE ECD ALIGNMENT

```
Homo S.    AUACUGUAAUUACCUUUGAGUUUA
Rattus N.  AUACUGUAAUUACCUUUGAGUUUA
Mus M.     AUACUGUAAUUACCUUUGAGUUUA
           |||||| |||||||||    ||||
Mus M.     UGUGACUUUAAUGGAGG---AAGU
Rattus N.  UGUGACUUUAAUGGAGG---AAGU
Homo S.    UGUGACUUUAAUGGAGG---AAGU
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Intronic**: ECD Half Separation=**1,915bp**: Where=**3 prime/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.24 Candidate Report for ENSMUSE00000395043

ENSEMBL GENE: **ENSMUSG00000031284**
DESCRIPTION: **Serine/threonine-protein p21-activated kinase 3 (PAK-3/Beta-PAK)**

FUNCTION: Highly expressed in postmitotic neurons of the developing and postnatal cerebral cortex and hippocampus. Defects in pak3 are the cause of non-specific x-linked nonsyndromic mental retardation type 30.

RANK: **24th** *LOD SCORE*: **18.3**
SPECIES DEPTH: [M] [R] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G.  AAGAAGC---UUGCAU-CAAAGUGCUUCU
Homo S.    AAGAAGC---UUGCAU-CAAAGUGCUUCU
Rattus N.  AAGAAGC---UUGCAU-CAAAGUGCUUCU
Mus M.     AAGAAGC---UUGCAU-CAAAGUGCUUCU
           |||||||   ||| || |||   ||||||
Mus M.     UUCUUCGUGAAAC-UACGUU---CGAAGA
Rattus N.  UUCUUCGUGAAAC-UACGUU---CGAAGA
Homo S.    UUCUUCGUGAAAC-UACGUU---CGAAGA
Gallus G.  UUCUUCGUGAAAC-UACGUU---CGAAGA
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

NOTE: This ECD is palindromic.

```
Homo S.    AAGAAGCUUGCAUCAAAGU-GCUUCU---UUGA
Rattus N.  AAGAAGCUUGCAUCAAAGU-GCUUCU---UUGA
Mus M.     AAGAAGCUUGCAUCAAAGU-GCUUCU---UUGA
           |||||||  |||| ||||| ||||||   ||||
Mus M.     UUUUUCG--CGUACUUUUAGUGGAGGUAUAACU
Rattus N.  UUUUUCG--UGUACUUUUAGUGGAGGUAUAACU
Homo S.    UUUUUCG--CGUACUUUUAGUGGAGGUAUAACU
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Intronic**: ECD Half Separation=**0bp**: Where=**3 prime/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.25 Candidate Report for ENSMUSE00000463801

ENSEMBL GENE: **ENSMUSG00000033565**
DESCRIPTION: **Fox-1 Homolog (C.Elegans). Hexaribonucleotide binding protein 2. Fyn-binding molecule. RNA binding motif protein 9.**
FUNCTION: Transcription factor with possible role in mRNA splicing.

RANK: **25th** *LOD SCORE*: **18.2**
SPECIES DEPTH: [M] [R] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G. CAG--GGAUUCGGGUUC-GUAACUUUCGAGAAUAG
Homo S.   CAG--GGAUUCGGGUUC-GUAACUUUCGAGAAUAG
Rattus N. CAG--GGAUUCGGGUUC-GUAACUUUCGAGAAUAG
Mus M.    CAG--GGAUUCGGGUUC-GUAACUUUCGAGAAUAG
          ||| ||||| |||||| |||||||| ||| |||
Mus M.    GUCGGUUUGACCUCAAGUCAUUGAGA-CUC--AUU
Rattus N. GUCGGUUUGACCUCAAGUCAUUGAGA-CUC--AUU
Homo S.   GUCGGUUUGACCUCAAGUCAUUGAGA-CUC--AUU
Gallus G. GUCGGUUUGACCUCAAGUCAUUGAGA-CUC--AUU
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Boundary**: ECD Half Separation=**1,198bp**: Where=**5 prime/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: (33rd) Novel ECD in same gene.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.26 Candidate Report for ENSMUSE00000099319

ENSEMBL GENE: **ENSMUSG00000019947**
DESCRIPTION: **AT-rich interactive domain-containing protein 5B (ARID domain-containing protein 5B) (Mrf1-like)**
FUNCTION: DNA-binding protein that binds to the 5'-AAUA[CU]-3' core sequence. Probably acts as a transcription regulator. Represses the cytomegalovirus enhancer. Overexpression leads to induction of smooth muscle marker genes, suggesting that it may act as a regulator of smooth muscle cell differentiation and proliferation. May be involved in lipid stores.

RANK: **26th** *LOD SCORE*: **18.2**
SPECIES DEPTH: [M] [R] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G. GAGGAUAUGAAACAGUAAGUGUUUA
Homo S.   GAGGAUAUGAAACAGUAAGUGUUUA
Rattus N. GAGGAUAUGAAACAGUAAGUGUUUA
Mus M.    GAGGAUAUGAAACAGUAAGUGUUUA
          *
Mus M.    CUCC----CU--GUUAUUCAUAAAU
Rattus N. CUCC----CU--GUUAUUCAUAAAU
Homo S.   CUCC----CU--GUUAUUCAUAAAU
Gallus G. CUCC----CU--GUUAUUCAUAAAU
```

Exonic Part (ES)

Intronic or 2nd Exonic Part (ECS)

ECS LOCATION Type=**Boundary**: ECD Half Separation=**53bp**: Where=**3 prime/5 prime**
The ECD halves flank the exon closely, overlapping both exon junctions.

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Figure A1.27 Candidate Report for ENSMUSE00000153249

ENSEMBL GENE: **ENSMUSG00000025789**
DESCRIPTION: **Alpha-2,8-sialyltransferase 8B**

FUNCTION: May transfer sialic acid through alpha-2,8-linkages to the alpha-2,3-linked and alpha-2,6-linked sialic acid of N-linked oligosaccharides of glycoproteins in the Golgi.

RANK: **27th** *LOD SCORE*: **18.2**
SPECIES DEPTH: [M] [R] [X] [G] [X] [X] & [M] [R] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G. UUGUCUUGCAGGAGUUCUGGAGGCAGA--GGUACAA  } Exonic Part
Rattus N. UUGUCUUGCAGGAAUUCUGGAGGCAGA--GGUACAA  } (ES)
Mus M.    UUGUCUUGCAGGAAUUCUGGAGGCAGA--GGUACAA  }
          ||||||||||       ||||||||| |||||||
Mus M.    GACGGAAUGUC------GUUUUUGUUUACUUGUGUU  } Intronic or
Rattus N. GACGGAAUGUC------GUUUUUGUUUACUUGUGUU  } 2nd Exonic
Gallus G. -ACGGAAUGUC------GUUUUUGUUUACUUGUGUU  } Part (ECS)
```

SECONDARY ECD ALIGNMENT (EXONIC)

```
Gallus G. ACAUAGCAAAUCUAAUAGGUUUGUAAAUU  } Exonic Part
Homo S.   ACAUAGCAAAUCUAAUAGGUUUGUAAAUU  } (ES)
Rattus N. ACAUAGCAAAUCUAAUAGGUUUGUAAAUU  }
Mus M.    ACAUAGCAAAUCUAAUAGGUUUGUAAAUU  }
          |||||  |||||||||    |||||||||
Mus M.    UGUGU--UUUAGAUUA----AAUGUUUGG  } Intronic or
Rattus N. UGUGU--UUUAGAUUA----AAUGUUUGG  } 2nd Exonic
Homo S.   UGUGU--UUUAGAUUA----AAUGUUUGG  } Part (ECS)
Gallus G. UGUAG--UUUA------CC--GUAUUUAA  }
```

ECS LOCATION Type=**Boundary**: ECD Half Separation=**76bp**: Where=**5 prime/3 prime**
The ECD halves flank the exon closely, overlapping both exon junctions.

ADDITIONAL INFORMATION
RELATED SITES: None.
A/G MISMATCHES: No A-G mismatch observed.

Figure A1.28 Candidate Report for ENSMUSE00000469222

ENSEMBL GENE: **ENSMUSG00000033981**
DESCRIPTION: **Glutamate receptor 2 precursor (GluR-2/GluR-B/GluR-K2/AMPA2)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **28th** *LOD SCORE*: **18.2**
SPECIES DEPTH: [M] [R] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G. UGCUCAC--CC-UGUCUGACAAGUA----UGUUUU  } Exonic Part
Homo S.   UGCUCAC--CC-UGUCUGACAAGUA----UGUUUU  } (ES)
Rattus N. UGCUCAC--CC-UGUCUGACAAGUA----UGUUUU  }
Mus M.    UGCUCAC--CC-UGUCUGACAAGUA----UGUUUU  }
          |||||||  || || | ||||||||    ||||||
Mus M.    AUGAGUGUAGGUGUUGUUUGUUUAUUGUCACAGAA  } Intronic or
Rattus N. AUGAGUGUAGGUGUUGUUUGUUUAUUGUCACAGAA  } 2nd Exonic
Homo S.   AUGAGUGUAGGUGUUGUUUGUUUAUUGUCACAGAA  } Part (ECS)
Gallus G. AUGAGUGUAGGUGUUGUUUGUUUAUUGUCACAGAA  }
```

SECONDARY ECS ALIGNMENT

```
Gallus G. AAGG----AGAGUGCGGCAGCGGGGGAGGUGA  } Exonic Part
Rattus N. AAGG----AGAGUGCGGCAGCGGGGGAGGUGA  } (ES)
Mus M.    AAGG----AGAGUGCGGCAGCGGGGGAGGUGA  }
            ||||    |||||| ||||| ||||||||||
Mus M.    UUCCACCAUCUCGCAUUGUC-UUUCUUUUACU  } Intronic or
Rattus N. UUCCACCAUCUCGUAUUGUU-UUUCUUUUACU  } 2nd Exonic
Gallus G. GAGGUG-AUUU-U-UUUGUUUGUUGUUUUAUC  } Part (ECS)
```

ECS LOCATION Type=**Standard**: ECD Half Separation=**392bp**: Where=**Exon/5 prime**

ADDITIONAL INFORMATION
RELATED SITES: (29th) Same exon (different 3'prime end)
(4th) GluR-B R/G Site
(5th) GluR-B Q/R Site
(2nd) GluR-B Novel Site
A/G MISMATCHES: A-G mismatch overlapping the ES.

Figure A1.29 Candidate Report for ENSMUSE00000487901

ENSEMBL GENE: **ENSMUSG00000033981**
DESCRIPTION: **Glutamate receptor 2 precursor (GluR-2/GluR-B/GluR-K2/AMPA2)**

FUNCTION: Glutamate-gated potassium channel involved in synaptic transmission.

RANK: **29th** *LOD SCORE*: **18.2**
SPECIES DEPTH: [M] [R] [H] [G] [X] [X]

CORE ECD ALIGNMENT

```
Gallus G. UGCUCAC--CC-UGUCUGACAAGUA----UGUUUU   }
Homo S.   UGCUCAC--CC-UGUCUGACAAGUA----UGUUUU   } Exonic Part
Rattus N. UGCUCAC--CC-UGUCUGACAAGUA----UGUUUU   } (ES)
Mus M.    UGCUCAC--CC-UGUCUGACAAGUA----UGUUUU   }
          |||||||  ||  ||  |  ||||||||    ||||||
Mus M.    AUGAGUGUAGGUGUUGUUUGUUUAUUGUCACAGAA   }
Rattus N. AUGAGUGUAGGUGUUGUUUGUUUAUUGUCACAGAA   } Intronic or
Homo S.   AUGAGUGUAGGUGUUGUUUGUUUAUUGUCACAGAA   } 2nd Exonic
Gallus G. AUGAGUGUAGGUGUUGUUUGUUUAUUGUCACAGAA   } Part (ECS)
```

SECONDARY ECS ALIGNMENT

```
Gallus G. AAGG----AGAGUGCGGCAGCGGGGGAGGUGA   }
Rattus N. AAGG----AGAGUGCGGCAGCGGGGGAGGUGA   } Exonic Part
Mus M.    AAGG----AGAGUGCGGCAGCGGGGGAGGUGA   } (ES)
           ||||    ||||||  |||||  ||||||||||
Mus M.    UUCCACCAUCUCGCAUUGUC-UUUCUUUUACU   }
Rattus N. UUCCACCAUCUCGUAUUGUU-UUUCUUUUACU   } Intronic or
Gallus G. GAGGUG-AUUU-U-UUUGUUUGUUGUUUUAUC   } 2nd Exonic
                                             } Part (ECS)
```

ECS LOCATION Type=**Standard**: ECD Half Separation=**392bp**: Where=**Exon/5 prime**

ADDITIONAL INFORMATION

RELATED SITES: (28th) Same exon (different 3'prime end)
(4th) GluR-B R/G Site
(5th) GluR-B Q/R Site
(2nd) GluR-B Novel Site

A/G MISMATCHES: A-G mismatch overlapping the ES.

Figure A1.30 Candidate Report for ENSMUSE00000178321

ENSEMBL GENE: **ENSMUSG00000061455**
DESCRIPTION: **Syntaxin variant**

FUNCTION: Synaptic activity. Implicated in vesicle trafficking to lysosomes.

RANK: **30th** *LOD SCORE*: **18.1**
SPECIES DEPTH: [M] [R] [X] [X] [D] [X]

CORE ECD ALIGNMENT

```
Danio R.  GUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGU   }
Rattus N. GUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGU   } Exonic Part
Mus M.    GUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGU   } (ES)
          ||||||||||||||||||||||||||||||||||
Mus M.    UGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUG   }
Rattus N. UGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUG   } Intronic or
Danio R.  UGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUU   } 2nd Exonic
                                             } Part (ECS)
```

NOTE: This ECD is both palindromic and composed of simple repeats.

PUTAUIVE RNA DUPLEX STRUCTURE

```
GUGUGUGUGUGUGU G U
               |||||||||||||    G
UGUGUGUGUGUGU U G               U
```

ECS LOCATION Type=**Intronic**: ECD Half Separation=**0bp**: Where=**3 prime/3 prime**

ADDITIONAL INFORMATION

RELATED SITES: None.

A/G MISMATCHES: No A-G mismatch observed.

Appendix 2. POCUS: Mining Genomic Sequence Annotation to Predict Disease Genes

Prior to this PhD I did two bioinformatic projects for my Masters degree, both of which were supervised by Semple lab. One of these projects was a pilot study for the prioritisation of disease genes using genomic sequence annotation. During the early stages of this PhD I collaborated with another student, Frances Turner, to develop this pilot study into a statistically rigorous algorithm that has now been published in Genome Biology. I was joint first author on this paper.

The basic concept of POCUS, the realisation of this protocol, is that the genes underlying any given disease are likely to have some similarities. These may be similar protein domains, similar functions or other shared features. Many complex and multigenic diseases have two or more regions of the human genome associated with them, but the underlying genes are not known. By observing the annotation of the genes in these regions, we could identify any over-represented InterPro domains or GO terms, which represent protein domains and gene functions, respectively. A complex statistical method was derived to score these shared terms.

POCUS provides high enrichment of real disease genes in the candidate gene shortlists it produces compared with the original large sets of positional candidates (up to 81-fold enrichment). In contrast to other existing methods, POCUS is able to suggest counterintuitive candidates.

The manuscript for this work is freely available on the Genome Biology website (http://genomebiology.com/2003/4/11/R75) or on the attached CD. The other authors' permissions have been obtained to include this.

Appendix 3: FANTOM3 Collaboration

Based on my experience with sequence analysis, I applied and was selected to join the FANTOM3 consortium. This is a group of collaborators who, in conjunction with the Riken Institute in Japan, have recently published a series of papers on the generation and analysis of large amounts of novel expressed sequence data and CAGE data[73].

My primary role in this large collaboration was to annotate a portion of these sequences based on a range of bioinformatic analyses. These included sequence conservation, repeat masking, CAGE alignments, existing EST/cDNA alignments, alignments to genomic sequences, and many others. Although this was nothing to do with RNA editing, this was a very interesting and valuable experience.

Appendix 4: Contents of Supplementary CD

The attached CD contains a variety of files that are relevant to this thesis. These include:

Online ECD Prediction Tool files

The Perl script and associated HTML, PHP, and CGI files are contained in a directory called 'Standalone'. These files would need minor modifications for use on any other system than the one it is currently installed on. This resource is described in Chapter 6.

POCUS Manuscript PDF

The PDF for the POCUS manuscript is attached with full permission from the authors (see Appendix 2).

Top 1000 Vertebrate ECD Predictions

This is a Microsoft Excel table containing details of the 1,000 top-scoring ECD predictions from the vertebrate screen in Chapter 4.

ECD Predictions in HTML Format

The 'html_ecss' folder contains dynamic HTML formatted ECD prediction alignments. The file 'ecs_results.html' links to a list of all the ECD predictions for the known edited sites and the top candidates from each screen described in Chapters 4 and 5. ***These files must be viewed in Internet Explorer.***

References

1. Keegan,L.P., Gallo,A., & O'Connell,M.A. The many roles of an RNA editor. *Nat. Rev. Genet.* **2**, 869-878 (2001).

2. Morse,D.P., Aruscavage,P.J., & Bass,B.L. RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. *Proc. Natl. Acad. Sci. U. S. A* **99**, 7906-7911 (2002).

3. Slavov,D., Crnogorac-Jurcevic,T., Clark,M., & Gardiner,K. Comparative analysis of the DRADA A-to-I RNA editing gene from mammals, pufferfish and zebrafish. *Gene* **250**, 53-60 (2000).

4. Tonkin,L.A., Saccomanno,L., Morse,D.P., Brodigan,T., Krause,M., & Bass,B.L. RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. *EMBO J.* **21**, 6025-6035 (2002).

5. Palladino,M.J., Keegan,L.P., O'Connell,M.A., & Reenan,R.A. dADAR, a Drosophila double-stranded RNA-specific adenosine deaminase is highly developmentally regulated and is itself a target for RNA editing. *RNA.* **6**, 1004-1018 (2000).

6. Schmitz-Linneweber,C., Regel,R., Du,T.G., Hupfer,H., Herrmann,R.G., & Maier,R.M. The plastid chromosome of Atropa belladonna and its comparison with that of Nicotiana tabacum: the role of RNA editing in generating divergence in the process of plant speciation. *Mol. Biol. Evol.* **19**, 1602-1612 (2002).

7. Pai,R.D., Oppegard,L.M., & Connell,G.J. Sequence and structural requirements for optimal guide RNA-directed insertional editing within Leishmania tarentolae. *RNA.* **9**, 469-483 (2003).

8. Wolf,J., Gerber,A.P., & Keller,W. tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. *EMBO J.* **21**, 3841-3851 (2002).

9. Basilio,C., WAHBA,A.J., LENGYEL,P., SPEYER,J.F., & OCHOA,S. Synthetic polynucleotides and the amino acid code. V. *Proc. Natl. Acad. Sci. U. S. A* **48**, 613-616 (1962).

10. Bass,B.L. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71**, 817-846 (2002).

11. Keegan,L.P., Leroy,A., Sproul,D., & O'Connell,M.A. Adenosine deaminases acting on RNA (ADARs): RNA-editing enzymes. *Genome Biol.* **5**, 209 (2004).

12. Schaub,M. & Keller,W. RNA editing by adenosine deaminases generates RNA and protein diversity. *Biochimie* **84**, 791-803 (2002).

13. Gerber,A., Grosjean,H., Melcher,T., & Keller,W. Tad1p, a yeast tRNA-specific adenosine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. *EMBO J.* **17**, 4780-4789 (1998).

14. Wagner,R.W., Yoo,C., Wrabetz,L., Kamholz,J., Buchhalter,J., Hassan,N.F., Khalili,K., Kim,S.U., Perussia,B., McMorris,F.A., & . Double-stranded RNA unwinding and modifying activity is detected ubiquitously in primary tissues and cell lines. *Mol. Cell Biol.* **10**, 5586-5590 (1990).

15. Melcher,T., Maas,S., Herb,A., Sprengel,R., Higuchi,M., & Seeburg,P.H. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *J. Biol. Chem.* **271**, 31795-31798 (1996).

16. Chen,C.X., Cho,D.S., Wang,Q., Lai,F., Carter,K.C., & Nishikura,K. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA.* **6**, 755-767 (2000).

17. Kallman,A.M., Sahlin,M., & Ohman,M. ADAR2 A-->I editing: site selectivity and editing efficiency are separate events. *Nucleic Acids Res.* **31**, 4874-4881 (2003).

18. Lehmann,K.A. & Bass,B.L. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry* **39**, 12875-12884 (2000).

19. Jaikaran,D.C., Collins,C.H., & MacMillan,A.M. Adenosine to inosine editing by ADAR2 requires formation of a ternary complex on the GluR-B R/G site. *J. Biol. Chem.* **277**, 37624-37629 (2002).

20. Cho,D.S., Yang,W., Lee,J.T., Shiekhattar,R., Murray,J.M., & Nishikura,K. Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. *J. Biol. Chem.* **278**, 17093-17102 (2003).

21. Yi-Brunozzi,H.Y., Stephens,O.M., & Beal,P.A. Conformational changes that occur during an RNA-editing adenosine deamination reaction. *J. Biol. Chem.* **276**, 37827-37833 (2001).

22. Rueter,S.M., Dawson,T.R., & Emeson,R.B. Regulation of alternative splicing by RNA editing. *Nature* **399**, 75-80 (1999).

23. Higuchi,M., Single,F.N., Kohler,M., Sommer,B., Sprengel,R., & Seeburg,P.H. RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell* **75**, 1361-1370 (1993).

24. Burns,C.M., Chu,H., Rueter,S.M., Hutchinson,L.K., Canton,H., Sanders-Bush,E., & Emeson,R.B. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* **387**, 303-308 (1997).

25. Gurevich,I., Tamir,H., Arango,V., Dwork,A.J., Mann,J.J., & Schmauss,C. Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron* **34**, 349-356 (2002).

26. Niswender,C.M., Herrick-Davis,K., Dilley,G.E., Meltzer,H.Y., Overholser,J.C., Stockmeier,C.A., Emeson,R.B., & Sanders-Bush,E. RNA editing of the human serotonin 5-HT2C receptor. alterations in suicide and implications for serotonergic pharmacotherapy. *Neuropsychopharmacology* **24**, 478-491 (2001).

27. Peixoto,A.A., Smith,L.A., & Hall,J.C. Genomic organization and evolution of alternative exons in a Drosophila calcium channel gene. *Genetics* **145**, 1003-1013 (1997).

28. Smith,L.A., Wang,X., Peixoto,A.A., Neumann,E.K., Hall,L.M., & Hall,J.C. A Drosophila calcium channel alpha1 subunit gene maps to a genetic locus associated with behavioral and visual defects. *J. Neurosci.* **16**, 7868-7879 (1996).

29. Bettler,B. & Mulle,C. Review: neurotransmitter receptors. II. AMPA and kainate receptors. *Neuropharmacology* **34**, 123-139 (1995).

30. Schmauss,C. & Howe,J.R. RNA editing of neurotransmitter receptors in the mammalian brain. *Sci. STKE.* **2002**, E26 (2002).

31. Hume,R.I., Dingledine,R., & Heinemann,S.F. Identification of a site in glutamate receptor subunits that controls calcium permeability. *Science* **253**, 1028-1031 (1991).

32. Lomeli,H., Mosbacher,J., Melcher,T., Hoger,T., Geiger,J.R., Kuner,T., Monyer,H., Higuchi,M., Bach,A., & Seeburg,P.H. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* **266**, 1709-1713 (1994).

33. Sommer,B., Keinanen,K., Verdoorn,T.A., Wisden,W., Burnashev,N., Herb,A., Kohler,M., Takagi,T., Sakmann,B., & Seeburg,P.H. Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. *Science* **249**, 1580-1585 (1990).

34. Palladino,M.J., Keegan,L.P., O'Connell,M.A., & Reenan,R.A. A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. *Cell* **102**, 437-449 (2000).

35. Higuchi,M., Maas,S., Single,F.N., Hartner,J., Rozov,A., Burnashev,N., Feldmeyer,D., Sprengel,R., & Seeburg,P.H. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406**, 78-81 (2000).

36. Wang,Q., Miyakoda,M., Yang,W., Khillan,J., Stachura,D.L., Weiss,M.J., & Nishikura,K. Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. *J. Biol. Chem.* **279**, 4952-4961 (2004).

37. Wang,Q., Khillan,J., Gadue,P., & Nishikura,K. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science* **290**, 1765-1768 (2000).

38. Miyamura,Y., Suzuki,T., Kono,M., Inagaki,K., Ito,S., Suzuki,N., & Tomita,Y. Mutations of the RNA-specific adenosine deaminase gene (DSRAD) are involved in dyschromatosis symmetrica hereditaria. *Am. J. Hum. Genet.* **73**, 693-699 (2003).

39. Kwak,S. & Kawahara,Y. Deficient RNA editing of GluR2 and neuronal death in amyotropic lateral sclerosis. *J. Mol. Med.* **83**, 110-120 (2005).

40. Brusa,R., Zimmermann,F., Koh,D.S., Feldmeyer,D., Gass,P., Seeburg,P.H., & Sprengel,R. Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. *Science* **270**, 1677-1680 (1995).

41. Greger,I.H., Khatri,L., Kong,X., & Ziff,E.B. AMPA receptor tetramerization is mediated by Q/R editing. *Neuron* **40**, 763-774 (2003).

42. Hartner,J.C., Schmittwolf,C., Kispert,A., Muller,A.M., Higuchi,M., & Seeburg,P.H. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. *J. Biol. Chem.* **279**, 4894-4902 (2004).

43. Chao,S.C., Lee,J.Y., Sheu,H.M., & Yang,M.H. A novel deletion mutation of the DSRAD gene in a Taiwanese patient with dyschromatosis symmetrica hereditaria. *Br. J. Dermatol.* **153**, 1064-1066 (2005).

44. Cui,Y., Wang,J., Yang,S., Gao,M., Chen,J.J., Yan,K.L., Xiao,F.L., Huang,W., & Zhang,X.J. Identification of a novel mutation in the DSRAD gene in a Chinese pedigree with dyschromatosis symmetrica hereditaria. *Arch. Dermatol. Res.* **296**, 543-545 (2005).

45. Li,M., Li,C., Hua,H., Zhu,W., Lu,Y., & Yang,L. Identification of two novel mutations in Chinese patients with Dyschromatosis symmetrica hereditaria. *Arch. Dermatol. Res.* 1-5 (2005).

46. Sun,X.K., Xu,A.E., Chen,J.F., & Tang,X. The double-RNA-specific adenosine deaminase (DSRAD) gene in dyschromatosis symmetrica hereditaria patients: two novel mutations and one previously described. *Br. J. Dermatol.* **153**, 342-345 (2005).

47. Zhang,X.J., He,P.P., Li,M., He,C.D., Yan,K.L., Cui,Y., Yang,S., Zhang,K.Y., Gao,M., Chen,J.J., Li,C.R., Jin,L., Chen,H.D., Xu,S.J., & Huang,W. Seven novel mutations of the ADAR gene in Chinese families and sporadic patients with dyschromatosis symmetrica hereditaria (DSH). *Hum. Mutat.* **23**, 629-630 (2004).

48. Revy,P., Muto,T., Levy,Y., Geissmann,F., Plebani,A., Sanal,O., Catalan,N., Forveille,M., Dufourcq-Labelouse,R., Gennery,A., Tezcan,I., Ersoy,F., Kayserili,H., Ugazio,A.G., Brousse,N., Muramatsu,M., Notarangelo,L.D., Kinoshita,K., Honjo,T., Fischer,A., & Durandy,A. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* **102**, 565-575 (2000).

49. Chester,A., Somasekaram,A., Tzimina,M., Jarmuz,A., Gisbourne,J., O'Keefe,R., Scott,J., & Navaratnam,N. The apolipoprotein B mRNA editing complex performs a multifunctional cycle and suppresses nonsense-mediated decay. *EMBO J.* **22**, 3971-3982 (2003).

50. Chester,A., Weinreb,V., Carter,C.W., Jr., & Navaratnam,N. Optimization of apolipoprotein B mRNA editing by APOBEC1 apoenzyme and the role of its auxiliary factor, ACF. *RNA.* **10**, 1399-1411 (2004).

51. Meier,J.C., Henneberger,C., Melnick,I., Racca,C., Harvey,R.J., Heinemann,U., Schmieden,V., & Grantyn,R. RNA editing produces glycine receptor alpha3(P185L), resulting in high agonist potency. *Nat. Neurosci.* (2005).

52. Maier,R.M., Zeltz,P., Kossel,H., Bonnard,G., Gualberto,J.M., & Grienenberger,J.M. RNA editing in plant mitochondria and chloroplasts. *Plant Mol. Biol.* **32**, 343-365 (1996).

53. Casey,J.L. RNA editing in hepatitis delta virus genotype III requires a branched double-hairpin RNA structure. *J. Virol.* **76**, 7385-7397 (2002).

54. Simpson,L., Sbicego,S., & Aphasizhev,R. Uridine insertion/deletion RNA editing in trypanosome mitochondria: a complex business. *RNA.* **9**, 265-276 (2003).

55. Bundschuh,R. Computational prediction of RNA editing sites. *Bioinformatics.* (2004).

56. Miller,D., Mahendran,R., Spottswood,M., Costandy,H., Wang,S., Ling,M.L., & Yang,N. Insertional editing in mitochondria of Physarum. *Semin. Cell Biol.* **4**, 261-266 (1993).

57. Steward,M., Vipond,I.B., Millar,N.S., & Emmerson,P.T. RNA editing in Newcastle disease virus. *J. Gen. Virol.* **74 (Pt 12)**, 2539-2547 (1993).

58. Villegas,J., Muller,I., Arredondo,J., Pinto,R., & Burzio,L.O. A putative RNA editing from U to C in a mouse mitochondrial transcript. *Nucleic Acids Res.* **30**, 1895-1901 (2002).

59. Nutt,S.L., Hoo,K.H., Rampersad,V., Deverill,R.M., Elliott,C.E., Fletcher,E.J., Adams,S.L., Korczak,B., Foldes,R.L., & Kamboj,R.K. Molecular characterization of the human EAA5 (GluR7) receptor: a high-affinity kainate receptor with novel potential RNA editing sites. *Receptors. Channels* **2**, 315-326 (1994).

60. Sharma,P.M., Bowman,M., Madden,S.L., Rauscher,F.J., III, & Sukumar,S. RNA editing in the Wilms' tumor susceptibility gene, WT1. *Genes Dev.* **8**, 720-731 (1994).

61. Novo,F.J., Kruszewski,A., MacDermot,K.D., Goldspink,G., & Gorecki,D.C. Editing of human alpha-galactosidase RNA resulting in a pyrimidine to purine conversion. *Nucleic Acids Res.* **23**, 2636-2640 (1995).

62. Liu,Z., Song,W., & Dong,K. Persistent tetrodotoxin-sensitive sodium current resulting from U-to-C RNA editing of an insect sodium channel. *Proc. Natl. Acad. Sci. U. S. A* **101**, 11862-11867 (2004).

63. Dawson,T.R., Sansam,C.L., & Emeson,R.B. Structure and sequence determinants required for the RNA editing of ADAR2 substrates. *J. Biol. Chem.* **279**, 4941-4951 (2004).

64. Zhang,Z. & Carmichael,G.G. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell* **106**, 465-475 (2001).

65. Peters,N.T., Rohrbach,J.A., Zalewski,B.A., Byrkett,C.M., & Vaughn,J.C. RNA editing and regulation of Drosophila 4f-rnp expression by sas-10 antisense readthrough mRNA transcripts. *RNA.* **9**, 698-710 (2003).

66. Scadden,A.D. & Smith,C.W. Specific cleavage of hyper-edited dsRNAs. *EMBO J.* **20**, 4243-4252 (2001).

67. Scadden,A.D. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. *Nat. Struct. Mol. Biol.* (2005).

68. Kim,D.D., Kim,T.T., Walsh,T., Kobayashi,Y., Matise,T.C., Buyske,S., & Gabriel,A. Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res.* **14**, 1719-1725 (2004).

69. Athanasiadis,A., Rich,A., & Maas,S. Widespread A-to-I RNA Editing of Alu-Containing mRNAs in the Human Transcriptome. *PLoS. Biol.* **2**, e391 (2004).

70. Levanon,E.Y., Eisenberg,E., Yelin,R., Nemzer,S., Hallegger,M., Shemesh,R., Fligelman,Z.Y., Shoshan,A., Pollock,S.R., Sztybel,D., Olshansky,M., Rechavi,G., & Jantsch,M.F. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **22**, 1001-1005 (2004).

71. Eisenberg,E., Nemzer,S., Kinar,Y., Sorek,R., Rechavi,G., & Levanon,E.Y. Is abundant A-to-I RNA editing primate-specific? *Trends Genet.* **21**, 77-81 (2005).

72. Chen,J., Sun,M., Hurst,L.D., Carmichael,G.G., & Rowley,J.D. Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends Genet.* **21**, 326-329 (2005).

73. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C., Kodzius,R., Shimokawa,K., Bajic,V.B., Brenner,S.E., Batalov,S., Forrest,A.R., Zavolan,M., Davis,M.J., Wilming,L.G., Aidinis,V., Allen,J.E., Ambesi-Impiombato,A., Apweiler,R., Aturaliya,R.N., Bailey,T.L., Bansal,M., Baxter,L., Beisel,K.W., Bersano,T., Bono,H., Chalk,A.M., Chiu,K.P., Choudhary,V., Christoffels,A., Clutterbuck,D.R., Crowe,M.L., Dalla,E., Dalrymple,B.P., de Bono,B., Della,G.G., di Bernardo,D., Down,T., Engstrom,P., Fagiolini,M., Faulkner,G., Fletcher,C.F., Fukushima,T., Furuno,M., Futaki,S., Gariboldi,M., Georgii-Hemming,P., Gingeras,T.R., Gojobori,T., Green,R.E., Gustincich,S., Harbers,M., Hayashi,Y., Hensch,T.K., Hirokawa,N., Hill,D., Huminiecki,L., Iacono,M., Ikeo,K., Iwama,A., Ishikawa,T., Jakt,M., Kanapin,A., Katoh,M., Kawasawa,Y., Kelso,J., Kitamura,H., Kitano,H., Kollias,G., Krishnan,S.P., Kruger,A., Kummerfeld,S.K., Kurochkin,I.V., Lareau,L.F., Lazarevic,D., Lipovich,L., Liu,J., Liuni,S., McWilliam,S., Madan,B.M., Madera,M., Marchionni,L., Matsuda,H., Matsuzawa,S., Miki,H., Mignone,F., Miyake,S., Morris,K., Mottagui-Tabar,S., Mulder,N., Nakano,N., Nakauchi,H., Ng,P., Nilsson,R., Nishiguchi,S., Nishikawa,S., Nori,F., Ohara,O., Okazaki,Y., Orlando,V., Pang,K.C., Pavan,W.J., Pavesi,G., Pesole,G., Petrovsky,N., Piazza,S., Reed,J., Reid,J.F., Ring,B.Z., Ringwald,M., Rost,B., Ruan,Y., Salzberg,S.L., Sandelin,A., Schneider,C., Schonbach,C., Sekiguchi,K., Semple,C.A., Seno,S., Sessa,L., Sheng,Y., Shibata,Y., Shimada,H., Shimada,K., Silva,D., Sinclair,B., Sperling,S., Stupka,E., Sugiura,K., Sultana,R., Takenaka,Y., Taki,K., Tammoja,K., Tan,S.L., Tang,S., Taylor,M.S., Tegner,J., Teichmann,S.A., Ueda,H.R., van Nimwegen,E., Verardo,R., Wei,C.L., Yagi,K., Yamanishi,H., Zabarovsky,E., Zhu,S., Zimmer,A., Hide,W., Bult,C., Grimmond,S.M., Teasdale,R.D., Liu,E.T., Brusic,V., Quackenbush,J., Wahlestedt,C., Mattick,J.S., Hume,D.A., Kai,C., Sasaki,D., Tomaru,Y., Fukuda,S., Kanamori-Katayama,M., Suzuki,M., Aoki,J., Arakawa,T., Iida,J., Imamura,K., Itoh,M., Kato,T., Kawaji,H., Kawagashira,N., Kawashima,T., Kojima,M., Kondo,S., Konno,H., Nakano,K., Ninomiya,N., Nishio,T., Okada,M., Plessy,C., Shibata,K., Shiraki,T., Suzuki,S., Tagami,M., Waki,K., Watahiki,A., Okamura-Oho,Y., Suzuki,H., Kawai,J., & Hayashizaki,Y. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563 (2005).

74. Cavaille,J., Buiting,K., Kiefmann,M., Lalande,M., Brannan,C.I., Horsthemke,B., Bachellerie,J.P., Brosius,J., & Huttenhofer,A. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl. Acad. Sci. U. S. A* **97**, 14311-14316 (2000).

75. Filipowicz,W. Imprinted expression of small nucleolar RNAs in brain: time for RNomics. *Proc. Natl. Acad. Sci. U. S. A* **97**, 14035-14037 (2000).

76. Yi-Brunozzi,H.Y., Easterwood,L.M., Kamilar,G.M., & Beal,P.A. Synthetic substrate analogs for the RNA-editing adenosine deaminase ADAR-2. *Nucleic Acids Res.* **27**, 2912-2917 (1999).

77. Lehmann,K.A. & Bass,B.L. The importance of internal loops within RNA substrates of ADAR1. *J. Mol. Biol.* **291**, 1-13 (1999).

78. Herb,A., Higuchi,M., Sprengel,R., & Seeburg,P.H. Q/R site editing in kainate receptor GluR5 and GluR6 pre-mRNAs requires distant intronic sequences. *Proc. Natl. Acad. Sci. U. S. A* **93**, 1875-1880 (1996).

79. Clutterbuck,D.R., Leroy,A., O'Connell,M.A., & Semple,C.A. A bioinformatic screen for novel A-I RNA editing sites reveals recoding editing in BC10. *Bioinformatics.* **21**, 2590-2595 (2005).

80. Bhalla,T., Rosenthal,J.J., Holmgren,M., & Reenan,R. Control of human potassium channel inactivation by editing of a small mRNA hairpin. *Nat. Struct. Mol. Biol.* **11**, 950-956 (2004).

81. Levanon,E.Y., Hallegger,M., Kinar,Y., Shemesh,R., Djinovic-Carugo,K., Rechavi,G., Jantsch,M.F., & Eisenberg,E. Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res.* **33**, 1162-1168 (2005).

82. Tanoue,A., Koshimizu,T.A., Tsuchiya,M., Ishii,K., Osawa,M., Saeki,M., & Tsujimoto,G. Two novel transcripts for human endothelin B receptor produced by RNA editing/alternative splicing from a single gene. *J. Biol. Chem.* **277**, 33205-33212 (2002).

83. Ma,J., Qian,R., Rausa,F.M., III, & Colley,K.J. Two naturally occurring alpha2,6-sialyltransferase forms with a single amino acid change in the catalytic domain differ in their catalytic activity and proteolytic processing. *J. Biol. Chem.* **272**, 672-679 (1997).

84. Seeburg,P.H. The TiPS/TINS lecture: the molecular biology of mammalian glutamate receptor channels. *Trends Pharmacol. Sci.* **14**, 297-303 (1993).

85. Aruscavage,P.J. & Bass,B.L. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. *RNA.* **6**, 257-269 (2000).

86. Hedges,S.B. The origin and evolution of model organisms. *Nat. Rev. Genet.* **3**, 838-849 (2002).

87. Woolfe,A., Goodson,M., Goode,D.K., Snell,P., McEwen,G.K., Vavouri,T., Smith,S.F., North,P., Callaway,H., Kelly,K., Walter,K., Abnizova,I., Gilks,W., Edwards,Y.J., Cooke,J.E., & Elgar,G. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS. Biol.* **3**, e7 (2005).

88. Glazov,E.A., Pheasant,M., McGraw,E.A., Bejerano,G., & Mattick,J.S. Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res.* (2005).

89. Ryter,J.M. & Schultz,S.C. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.* **17**, 7505-7513 (1998).

90. Ramos,A., Grunert,S., Adams,J., Micklem,D.R., Proctor,M.R., Freund,S., Bycroft,M., St Johnston,D., & Varani,G. RNA recognition by a Staufen double-stranded RNA-binding domain. *EMBO J.* **19**, 997-1009 (2000).

91. Stephens,O.M., Haudenschild,B.L., & Beal,P.A. The binding selectivity of ADAR2's dsRBMs contributes to RNA-editing selectivity. *Chem. Biol.* **11**, 1239-1250 (2004).

92. Wong,S.K., Sato,S., & Lazinski,D.W. Substrate recognition by ADAR1 and ADAR2. *RNA.* **7**, 846-858 (2001).

93. Herbert,A., Lowenhaupt,K., Spitzner,J., & Rich,A. Chicken double-stranded RNA adenosine deaminase has apparent specificity for Z-DNA. *Proc. Natl. Acad. Sci. U. S. A* **92**, 7550-7554 (1995).

94. Herbert,A. RNA editing, introns and evolution. *Trends Genet.* **12**, 6-9 (1996).

95. Herbert,A., Alfken,J., Kim,Y.G., Mian,I.S., Nishikura,K., & Rich,A. A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase. *Proc. Natl. Acad. Sci. U. S. A* **94**, 8421-8426 (1997).

96. Schwartz,T., Rould,M.A., Lowenhaupt,K., Herbert,A., & Rich,A. Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science* **284**, 1841-1845 (1999).

97. Davidson,N.O. The challenge of target sequence specificity in C-->U RNA editing. *J. Clin. Invest* **109**, 291-294 (2002).

98. Skuse,G.R., Cappione,A.J., Sowden,M., Metheny,L.J., & Smith,H.C. The neurofibromatosis type I messenger RNA undergoes base-modification RNA editing. *Nucleic Acids Res.* **24**, 478-485 (1996).

99. Patterson,J.B. & Samuel,C.E. Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Mol. Cell Biol.* **15**, 5376-5388 (1995).

100. Yang,J.H., Luo,X., Nie,Y., Su,Y., Zhao,Q., Kabir,K., Zhang,D., & Rabinovici,R. Widespread inosine-containing mRNA in lymphocytes regulated by ADAR1 in response to inflammation. *Immunology* **109**, 15-23 (2003).

101. Knight,S.W. & Bass,B.L. The role of RNA editing by ADARs in RNAi. *Mol. Cell* **10**, 809-817 (2002).

102. Tonkin,L.A. & Bass,B.L. Mutations in RNAi rescue aberrant chemotaxis of ADAR mutants. *Science* **302**, 1725 (2003).

103. Scadden,A.D. & Smith,C.W. RNAi is antagonized by A-->I hyper-editing. *EMBO Rep.* **2**, 1107-1111 (2001).

104. Yang,W., Wang,Q., Howell,K.L., Lee,J.T., Cho,D.S., Murray,J.M., & Nishikura,K. ADAR1 RNA deaminase limits short interfering RNA efficacy in mammalian cells. *J. Biol. Chem.* **280**, 3946-3953 (2005).

105. DeCerbo,J. & Carmichael,G.G. Retention and repression: fates of hyperedited RNAs in the nucleus. *Curr. Opin. Cell Biol.* **17**, 302-308 (2005).

106. Polson,A.G., Ley,H.L., III, Bass,B.L., & Casey,J.L. Hepatitis delta virus RNA editing is highly specific for the amber/W site and is suppressed by hepatitis delta antigen. *Mol. Cell Biol.* **18**, 1919-1926 (1998).

107. Bishop,K.N., Holmes,R.K., Sheehy,A.M., & Malim,M.H. APOBEC-mediated editing of viral RNA. *Science* **305**, 645 (2004).

108. Reenan,R.A. Molecular determinants and guided evolution of species-specific RNA editing. *Nature* **434**, 409-413 (2005).

109. Wang,Q., Zhang,Z., Blackwell,K., & Carmichael,G.G. Vigilins bind to promiscuously A-to-I-edited RNAs and are involved in the formation of heterochromatin. *Curr. Biol.* **15**, 384-391 (2005).

110. Paul,M.S. & Bass,B.L. Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *EMBO J.* **17**, 1120-1127 (1998).

111. Beghini,A., Ripamonti,C.B., Peterlongo,P., Roversi,G., Cairoli,R., Morra,E., & Larizza,L. RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. *Hum. Mol. Genet.* **9**, 2297-2304 (2000).

112. Bratt,E. & Ohman,M. Coordination of editing and splicing of glutamate receptor pre-mRNA. *RNA.* **9**, 309-318 (2003).

113. Reenan,R.A., Hanrahan,C.J., & Barry,G. The mle(napts) RNA helicase mutation in drosophila results in a splicing catastrophe of the para Na+ channel transcript in a region of RNA editing. *Neuron* **25**, 139-149 (2000).

114. Lee,C.G. & Hurwitz,J. A new RNA helicase isolated from HeLa cells that catalytically translocates in the 3' to 5' direction. *J. Biol. Chem.* **267**, 4398-4407 (1992).

115. Raitskin,O., Cho,D.S., Sperling,J., Nishikura,K., & Sperling,R. RNA editing activity is associated with splicing factors in lnRNP particles: The nuclear pre-mRNA processing machinery. *Proc. Natl. Acad. Sci. U. S. A* **98**, 6571-6576 (2001).

116. Flomen,R., Knight,J., Sham,P., Kerwin,R., & Makoff,A. Evidence that RNA editing modulates splice site selection in the 5-HT2C receptor gene. *Nucleic Acids Res.* **32**, 2113-2122 (2004).

117. Agrawal,R. & Stormo,G.D. Editing efficiency of a Drosophila gene correlates with a distant splice site selection. *RNA.* **11**, 563-566 (2005).

118. Lian,Y. & Garner,H.R. Evidence for the regulation of alternative splicing via complementary DNA sequence repeats. *Bioinformatics.* **21**, 1358-1364 (2005).

119. Miriami,E., Margalit,H., & Sperling,R. Conserved sequence elements associated with exon skipping. *Nucleic Acids Res.* **31**, 1974-1983 (2003).

120. Graveley,B.R. Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell* **123**, 65-73 (2005).

121. Grover,A., Houlden,H., Baker,M., Adamson,J., Lewis,J., Prihar,G., Pickering-Brown,S., Duff,K., & Hutton,M. 5' splice site mutations in tau associated with the inherited dementia FTDP-17 affect a stem-loop structure that regulates alternative splicing of exon 10. *J. Biol. Chem.* **274**, 15134-15143 (1999).

122. Blanchette,M. & Chabot,B. A highly stable duplex structure sequesters the 5' splice site region of hnRNP A1 alternative exon 7B. *RNA.* **3**, 405-419 (1997).

123. Cohen,R.S., Zhang,S., & Dollar,G.L. The positional, structural, and sequence requirements of the Drosophila TLS RNA localization element. *RNA.* **11**, 1017-1029 (2005).

124. Kortenbruck,G., Berger,E., Speckmann,E.J., & Musshoff,U. RNA editing at the Q/R site for the glutamate receptor subunits GLUR2, GLUR5, and GLUR6 in hippocampus and temporal cortex from epileptic patients. *Neurobiol. Dis.* **8**, 459-468 (2001).

125. Vollmar,W., Gloger,J., Berger,E., Kortenbruck,G., Kohling,R., Speckmann,E.J., & Musshoff,U. RNA editing (R/G site) and flip-flop splicing of the AMPA receptor subunit GluR2 in nervous tissue of epilepsy patients. *Neurobiol. Dis.* **15**, 371-379 (2004).

126. Sodhi,M.S., Burnet,P.W., Makoff,A.J., Kerwin,R.W., & Harrison,P.J. RNA editing of the 5-HT(2C) receptor is reduced in schizophrenia. *Mol. Psychiatry* **6**, 373-379 (2001).

127. Iwamoto,K. & Kato,T. RNA editing of serotonin 2C receptor in human postmortem brains of major mental disorders. *Neurosci. Lett.* **346**, 169-172 (2003).

128. Akbarian,S., Smith,M.A., & Jones,E.G. Editing for an AMPA receptor subunit RNA in prefrontal cortex and striatum in Alzheimer's disease, Huntington's disease and schizophrenia. *Brain Res.* **699**, 297-304 (1995).

129. Maas,S., Patt,S., Schrey,M., & Rich,A. Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. *Proc. Natl. Acad. Sci. U. S. A* **98**, 14687-14692 (2001).

130. Cattaneo,R., Schmid,A., Eschle,D., Baczko,K., ter,M., V, & Billeter,M.A. Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. *Cell* **55**, 255-265 (1988).

131. Morse,D.P. & Bass,B.L. Detection of inosine in messenger RNA by inosine-specific cleavage. *Biochemistry* **15**, 8429-8434 (1997).

132. Luciano,D.J., Mirsky,H., Vendetti,N.J., & Maas,S. RNA editing of a miRNA precursor. *RNA.* **10**, 1174-1177 (2004).

133. Xia,S., Yang,J., Su,Y., Qian,J., Ma,E., & Haddad,G.G. Identification of new targets of Drosophila pre-mRNA adenosine deaminase. *Physiol Genomics* **20**, 195-202 (2005).

134. Kikuno,R., Nagase,T., Waki,M., & Ohara,O. HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.* **30**, 166-168 (2002).

135. Stapleton,M., Carlson,J., Brokstein,P., Yu,C., Champe,M., George,R., Guarin,H., Kronmiller,B., Pacleb,J., Park,S., Wan,K., Rubin,G.M., & Celniker,S.E. A Drosophila full-length cDNA resource. *Genome Biol.* **3**, RESEARCH0080 (2002).

136. Muirhead,C.A., Glass,N.L., & Slatkin,M. Multilocus self-recognition systems in fungi as a cause of trans-species polymorphism. *Genetics* **161**, 633-641 (2002).

137. Blow,M., Futreal,P.A., Wooster,R., & Stratton,M.R. A survey of RNA editing in human brain. *Genome Res.* **14**, 2379-2387 (2004).

138. Eisenberg,E., Adamsky,K., Cohen,L., Amariglio,N., Hirshberg,A., Rechavi,G., & Levanon,E.Y. Identification of RNA editing sites in the SNP database. *Nucleic Acids Res.* **33**, 4612-4617 (2005).

139. Hanrahan,C.J., Palladino,M.J., Ganetzky,B., & Reenan,R.A. RNA editing of the Drosophila para Na(+) channel transcript. Evolutionary conservation and developmental regulation. *Genetics* **155**, 1149-1160 (2000).

140. Hoopengardner,B., Bhalla,T., Staber,C., & Reenan,R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301**, 832-836 (2003).

141. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R., Gage,D., Harris,K., Heaford,A., Howland,J., Kann,L., Lehoczky,J., Levine,R., McEwan,P., McKernan,K., Meldrim,J., Mesirov,J.P., Miranda,C., Morris,W., Naylor,J., Raymond,C., Rosetti,M., Santos,R., Sheridan,A., Sougnez,C., Stange-Thomann,N., Stojanovic,N., Subramanian,A., Wyman,D., Rogers,J., Sulston,J., Ainscough,R., Beck,S., Bentley,D., Burton,J., Clee,C., Carter,N., Coulson,A., Deadman,R., Deloukas,P., Dunham,A., Dunham,I., Durbin,R., French,L., Grafham,D., Gregory,S., Hubbard,T., Humphray,S., Hunt,A., Jones,M., Lloyd,C., McMurray,A., Matthews,L., Mercer,S., Milne,S., Mullikin,J.C., Mungall,A., Plumb,R., Ross,M., Shownkeen,R., Sims,S., Waterston,R.H., Wilson,R.K., Hillier,L.W., McPherson,J.D., Marra,M.A., Mardis,E.R., Fulton,L.A., Chinwalla,A.T., Pepin,K.H., Gish,W.R., Chissoe,S.L., Wendl,M.C., Delehaunty,K.D., Miner,T.L., Delehaunty,A., Kramer,J.B., Cook,L.L., Fulton,R.S., Johnson,D.L., Minx,P.J., Clifton,S.W., Hawkins,T., Branscomb,E., Predki,P., Richardson,P., Wenning,S., Slezak,T., Doggett,N., Cheng,J.F., Olsen,A., Lucas,S., Elkin,C., Uberbacher,E., Frazier,M., Gibbs,R.A., Muzny,D.M., Scherer,S.E., Bouck,J.B., Sodergren,E.J., Worley,K.C., Rives,C.M., Gorrell,J.H., Metzker,M.L., Naylor,S.L., Kucherlapati,R.S., Nelson,D.L., Weinstock,G.M., Sakaki,Y., Fujiyama,A., Hattori,M., Yada,T., Toyoda,A., Itoh,T., Kawagoe,C., Watanabe,H., Totoki,Y., Taylor,T., Weissenbach,J., Heilig,R., Saurin,W., Artiguenave,F., Brottier,P., Bruls,T., Pelletier,E., Robert,C., Wincker,P., Smith,D.R., Doucette-Stamm,L., Rubenfield,M., Weinstock,K., Lee,H.M., Dubois,J., Rosenthal,A., Platzer,M., Nyakatura,G., Taudien,S., Rump,A., Yang,H., Yu,J., Wang,J., Huang,G., Gu,J., Hood,L., Rowen,L., Madan,A., Qin,S., Davis,R.W., Federspiel,N.A., Abola,A.P., Proctor,M.J., Myers,R.M., Schmutz,J., Dickson,M., Grimwood,J., Cox,D.R., Olson,M.V., Kaul,R., Raymond,C., Shimizu,N., Kawasaki,K., Minoshima,S., Evans,G.A., Athanasiou,M., Schultz,R., Roe,B.A., Chen,F., Pan,H., Ramser,J., Lehrach,H., Reinhardt,R., McCombie,W.R., de la,B.M., Dedhia,N., Blocker,H., Hornischer,K., Nordsiek,G., Agarwala,R., Aravind,L., Bailey,J.A., Bateman,A., Batzoglou,S., Birney,E., Bork,P., Brown,D.G., Burge,C.B., Cerutti,L., Chen,H.C., Church,D., Clamp,M., Copley,R.R., Doerks,T., Eddy,S.R., Eichler,E.E., Furey,T.S., Galagan,J., Gilbert,J.G., Harmon,C., Hayashizaki,Y., Haussler,D., Hermjakob,H., Hokamp,K., Jang,W., Johnson,L.S., Jones,T.A., Kasif,S., Kaspryzk,A., Kennedy,S., Kent,W.J., Kitts,P., Koonin,E.V., Korf,I., Kulp,D., Lancet,D., Lowe,T.M., McLysaght,A., Mikkelsen,T., Moran,J.V., Mulder,N., Pollara,V.J., Ponting,C.P., Schuler,G., Schultz,J., Slater,G., Smit,A.F., Stupka,E., Szustakowski,J., Thierry-Mieg,D., Thierry-Mieg,J., Wagner,L., Wallis,J., Wheeler,R., Williams,A., Wolf,Y.I., Wolfe,K.H., Yang,S.P., Yeh,R.F., Collins,F., Guyer,M.S., Peterson,J., Felsenfeld,A., Wetterstrand,K.A., Patrinos,A., Morgan,M.J., de Jong,P., Catanese,J.J., Osoegawa,K., Shizuya,H., Choi,S., & Chen,Y.J. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

142. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M., Hubbard,T., Kasprzyk,A., Keefe,D., Lehvaslaiho,H., Iyer,V., Melsopp,C., Mongin,E., Pettett,R., Potter,S., Rust,A., Schmidt,E., Searle,S., Slater,G., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Stupka,E., Ureta-Vidal,A., Vastrik,I., & Birney,E. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38-42 (2003).

143. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M., & Sirotkin,K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308-311 (2001).

144. Velculescu,V.E., Zhang,L., Vogelstein,B., & Kinzler,K.W. Serial analysis of gene expression. *Science* **270**, 484-487 (1995).

145. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T., Fukuda,S., Sasaki,D., Podhajska,A., Harbers,M., Kawai,J., Carninci,P., & Hayashizaki,Y. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A* **100**, 15776-15781 (2003).

146. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F., & . Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-1656 (1991).

147. Rawlings,C.J. & Searls,D.B. Computational gene discovery and human disease. *Curr. Opin. Genet. Dev.* **7**, 416-423 (1997).

148. Sorek,R., Shamir,R., & Ast,G. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20**, 68-71 (2004).

149. Drysdale,R.A. & Crosby,M.A. FlyBase: genes and gene models. *Nucleic Acids Res.* **33**, D390-D395 (2005).

150. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A., & Wheeler,D.L. GenBank. *Nucleic Acids Res.* **27**, 12-17 (1999).

151. Boguski,M.S., Lowe,T.M., & Tolstoshev,C.M. dbEST--database for "expressed sequence tags". *Nat. Genet.* **4**, 332-333 (1993).

152. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P., Antonarakis,S.E., Attwood,J., Baertsch,R., Bailey,J., Barlow,K., Beck,S., Berry,E., Birren,B., Bloom,T., Bork,P., Botcherby,M., Bray,N., Brent,M.R., Brown,D.G., Brown,S.D., Bult,C., Burton,J., Butler,J., Campbell,R.D., Carninci,P., Cawley,S., Chiaromonte,F., Chinwalla,A.T., Church,D.M., Clamp,M., Clee,C., Collins,F.S., Cook,L.L., Copley,R.R., Coulson,A., Couronne,O., Cuff,J., Curwen,V., Cutts,T., Daly,M., David,R., Davies,J., Delehaunty,K.D., Deri,J., Dermitzakis,E.T., Dewey,C., Dickens,N.J., Diekhans,M., Dodge,S., Dubchak,I., Dunn,D.M., Eddy,S.R., Elnitski,L., Emes,R.D., Eswara,P., Eyras,E., Felsenfeld,A., Fewell,G.A., Flicek,P., Foley,K., Frankel,W.N., Fulton,L.A., Fulton,R.S., Furey,T.S., Gage,D., Gibbs,R.A., Glusman,G., Gnerre,S., Goldman,N., Goodstadt,L., Grafham,D., Graves,T.A., Green,E.D., Gregory,S., Guigo,R., Guyer,M., Hardison,R.C., Haussler,D., Hayashizaki,Y., Hillier,L.W., Hinrichs,A., Hlavina,W., Holzer,T., Hsu,F., Hua,A., Hubbard,T., Hunt,A., Jackson,I., Jaffe,D.B., Johnson,L.S., Jones,M., Jones,T.A., Joy,A., Kamal,M., Karlsson,E.K., Karolchik,D., Kasprzyk,A., Kawai,J., Keibler,E., Kells,C., Kent,W.J., Kirby,A., Kolbe,D.L., Korf,I., Kucherlapati,R.S., Kulbokas,E.J., Kulp,D., Landers,T., Leger,J.P., Leonard,S., Letunic,I., Levine,R., Li,J., Li,M., Lloyd,C., Lucas,S., Ma,B., Maglott,D.R., Mardis,E.R., Matthews,L., Mauceli,E., Mayer,J.H., McCarthy,M., McCombie,W.R., McLaren,S., McLay,K., McPherson,J.D., Meldrim,J., Meredith,B., Mesirov,J.P., Miller,W., Miner,T.L., Mongin,E., Montgomery,K.T., Morgan,M., Mott,R., Mullikin,J.C., Muzny,D.M., Nash,W.E., Nelson,J.O., Nhan,M.N., Nicol,R., Ning,Z., Nusbaum,C., O'Connor,M.J., Okazaki,Y., Oliver,K., Overton-Larty,E., Pachter,L., Parra,G., Pepin,K.H., Peterson,J., Pevzner,P., Plumb,R., Pohl,C.S., Poliakov,A., Ponce,T.C., Ponting,C.P., Potter,S., Quail,M., Reymond,A., Roe,B.A., Roskin,K.M., Rubin,E.M., Rust,A.G., Santos,R., Sapojnikov,V., Schultz,B., Schultz,J., Schwartz,M.S., Schwartz,S., Scott,C., Seaman,S., Searle,S., Sharpe,T., Sheridan,A., Shownkeen,R., Sims,S., Singer,J.B., Slater,G., Smit,A., Smith,D.R., Spencer,B., Stabenau,A., Stange-Thomann,N., Sugnet,C., Suyama,M., Tesler,G., Thompson,J., Torrents,D., Trevaskis,E., Tromp,J., Ucla,C., Ureta-Vidal,A., Vinson,J.P., Von Niederhausern,A.C., Wade,C.M., Wall,M., Weber,R.J., Weiss,R.B., Wendl,M.C., West,A.P., Wetterstrand,K., Wheeler,R., Whelan,S., Wierzbowski,J., Willey,D., Williams,S., Wilson,R.K., Winter,E., Worley,K.C., Wyman,D., Yang,S., Yang,S.P., Zdobnov,E.M., Zody,M.C., & Lander,E.S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

153. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., & Lipman,D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

154. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A., & Batzoglou,S. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721-731 (2003).

155. Needleman,S.B. & Wunsch,C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453 (1970).

156. Pollard,D.A., Bergman,C.M., Stoye,J., Celniker,S.E., & Eisen,M.B. Benchmarking tools for the alignment of functional noncoding DNA. *BMC. Bioinformatics.* **5**, 6 (2004).

157. Rice,P., Longden,I., & Bleasby,A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).

158. Pearson,W.R. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132**, 185-219 (2000).

159. Smith,T.F., Waterman,M.S., & Fitch,W.M. Comparative biosequence metrics. *J. Mol. Evol.* **18**, 38-46 (1981).

160. Smit, AFA, Hubley, R, and Green, P. RepeatMasker at http://www.repeatmasker.org. 1-1-1996.
Ref Type: Unpublished Work

161. Otto,S.P. & Yong,P. The evolution of gene duplicates. *Adv. Genet.* **46**, 451-483 (2002).

162. Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B., & Bartel,D.P. The microRNAs of Caenorhabditis elegans. *Genes Dev.* **17**, 991-1008 (2003).

163. Furey,T.S., Diekhans,M., Lu,Y., Graves,T.A., Oddy,L., Randall-Maher,J., Hillier,L.W., Wilson,R.K., & Haussler,D. Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing. *Genome Res.* **14**, 2034-2040 (2004).

164. Burset,M., Seledtsov,I.A., & Solovyev,V.V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364-4375 (2000).

165. Mott,R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13**, 477-478 (1997).

166. Rost,B. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**, 525-539 (1996).

167. Varani,G. & McClain,W.H. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep.* **1**, 18-23 (2000).

168. Emeson,R.B. & Singh,M. RNA Editing. (2000).

169. Sommer,B., Kohler,M., Sprengel,R., & Seeburg,P.H. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11-19 (1991).

170. Gromova,I., Gromov,P., & Celis,J.E. bc10: A novel human bladder cancer-associated protein with a conserved genomic structure downregulated in invasive cancer. *Int. J. Cancer* **98**, 539-546 (2002).

171. Rae,F.K., Stephenson,S.A., Nicol,D.L., & Clements,J.A. Novel association of a diverse range of genes with renal cell carcinoma as identified by differential display. *Int. J. Cancer* **88**, 726-732 (2000).

172. Yamanaka,S., Balestra,M.E., Ferrell,L.D., Fan,J., Arnold,K.S., Taylor,S., Taylor,J.M., & Innerarity,T.L. Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. *Proc. Natl. Acad. Sci. U. S. A* **92**, 8483-8487 (1995).

173. Yamanaka,S., Poksay,K.S., Arnold,K.S., & Innerarity,T.L. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. *Genes Dev.* **11**, 321-333 (1997).

174. Zuker,M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406-3415 (2003).

175. Eddy,S.R. How do RNA folding algorithms work? *Nat. Biotechnol.* **22**, 1457-1458 (2004).

176. Reed,R. & Maniatis,T. The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes Dev.* **2**, 1268-1276 (1988).

177. Koike,M., Tsukada,S., Tsuzuki,K., Kijima,H., & Ozawa,S. Regulation of kinetic properties of GluR2 AMPA receptor channels by alternative splicing. *J. Neurosci.* **20**, 2166-2174 (2000).

178. Eichler,E.E. & Sankoff,D. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**, 793-797 (2003).

179. Berriz,G.F., King,O.D., Bryant,B., Sander,C., & Roth,F.P. Characterizing gene sets with FuncAssociate. *Bioinformatics.* **19**, 2502-2504 (2003).

180. Eckmann,C.R., Neunteufl,A., Pfaffstetter,L., & Jantsch,M.F. The human but not the Xenopus RNA-editing enzyme ADAR1 has an atypical nuclear localization signal and displays the characteristics of a shuttling protein. *Mol. Biol. Cell* **12**, 1911-1924 (2001).

181. Nie,Y., Zhao,Q., Su,Y., & Yang,J.H. Subcellular distribution of ADAR1 isoforms is synergistically determined by three nuclear discrimination signals and a regulatory motif. *J. Biol. Chem.* **279**, 13249-13255 (2004).

182. Keegan,L.P., Brindle,J., Gallo,A., Leroy,A., Reenan,R.A., & O'Connell,M.A. Tuning of RNA editing by ADAR is required in Drosophila. *EMBO J.* **24**, 2183-2193 (2005).

183. Smith,L.A., Peixoto,A.A., & Hall,J.C. RNA editing in the Drosophila DMCA1A calcium-channel alpha 1 subunit transcript. *J. Neurogenet.* **12**, 227-240 (1998).

184. Hanrahan,C.J., Palladino,M.J., Bonneau,L.J., & Reenan,R.A. RNA editing of a Drosophila sodium channel gene. *Ann. N. Y. Acad. Sci.* **868**, 51-66 (1999).

185. Semenov,E.P. & Pak,W.L. Diversification of Drosophila chloride channel gene by multiple posttranscriptional mRNA modifications. *J. Neurochem.* **72**, 66-72 (1999).

186. Grauso,M., Reenan,R.A., Culetto,E., & Sattelle,D.B. Novel putative nicotinic acetylcholine receptor subunit genes, Dalpha5, Dalpha6 and Dalpha7, in Drosophila melanogaster identify a new and highly conserved target of adenosine deaminase acting on RNA-mediated A-to-I pre-mRNA editing. *Genetics* **160**, 1519-1533 (2002).

187. Kawahara,Y., Ito,K., Sun,H., Aizawa,H., Kanazawa,I., & Kwak,S. Glutamate receptors: RNA editing and death of motor neurons. *Nature* **427**, 801 (2004).

188. Bailey,T.L. & Elkan,C. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 21-29 (1995).

189. Morse,D.P. & Bass,B.L. Long RNA hairpins that contain inosine are present in Caenorhabditis elegans poly(A)+ RNA. *Proc. Natl. Acad. Sci. U. S. A* **96**, 6048-6053 (1999).

190. Kung,S.S., Chen,Y.C., Lin,W.H., Chen,C.C., & Chow,W.Y. Q/R RNA editing of the AMPA receptor subunit 2 (GRIA2) transcript evolves no later than the appearance of cartilaginous fishes. *FEBS Lett.* **509**, 277-281 (2001).

Evolution is cleverer than you are.
- Leslie Orgel (Orgel's Second Rule)